Promoter and Enhancer Chromatin Dynamics during Direct Cell Fate Programming

Dissertation
zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)
im Fach Biologie

eingereicht an der
Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von
Mahmoud Ibrahim

Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr. Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät
Prof. Dr. Bernhard Grimm

Gutachter/innen:
1- Prof. Dr. Uwe Ohler
2- Prof. Dr. Esteban O. Mazzoni
3- Prof. Dr. Martin Vingron
Tag der mündlichen Prüfung: 17 Februar 2017

Contents

List of Figures

Abstract

Die Beschreibung genregulatorischer Ereignisse, die Veränderungen in der Genexpression bewirken, ist entscheidend um Zelldifferenzierung und -entwicklung zu verstehen. Dynamische Vernderungen der Chromatinstruktur, Histonmodifikationen und das Binden von Transkriptionsfaktoren an Enhancer und Promotoren, können mit Hilfe von genomweiten Hochdurchsatz-Sequenziertechniken wie ChIP-Seq, DNase-Seq, ATAC-Seq und RNA-Seq untersucht werden.

In dieser Arbeit entwickele ich mehrere probabilistische Modelle für die Analyse von genomweiten Sequenzierungsdaten. Diese umfassen 1. einen Peak-Finder für ChIP-/DNase-/ATAC-Seq-Daten, der sich Replikate zunutze macht und präzise Peak-Weiten berechnet, 2. eine Pipeline, die ein Hidden-Markov-Modell nutzt, um das Genom in hoher Auflösung in eindeutige Klassen von Kombinationen von Histonmodifikationen zu segmentieren, 3. ein Bayes-Netzwerk-Modell, welches multiple, zeitlich aufgelöste Histonmodifikations-ChIP-seq-Daten kombinatorisch clustert, um, basierend auf der Chromatinstruktur, Klassen von regulatorischen Elementen zu identifizieren.

Mit Hilfe dieser Modelle untersuchen wir die Promotorumgeben und zeigen einen Zusammenhang zwischen Chromatinstruktur und Promotordirektionalität. Darüber hinaus verwenden wir ein Modell zur direkten Reprogrammierung von Stammzellen in Motorneuronen durch die gezielte Expression von Transkriptionsfaktoren und analysieren die dadurch induzierten zeitlichen Vernderungen der Chromatinstruktur und Transkriptionsfaktorbindedynamik. Wir beobachten, dass Promotoren verschiedenen Chromatin-Dynamiken zur Aktivierung und Repression folgen, die mit den Chromatin-Dynamiken von Enhancer-Elementen korrelieren. Enhancer hingegen werden durch kooperatives Verhalten direkt induzierter Transkriptionsfaktoren und anderen Faktoren, die in den Stammzellen zu Beginn vorhanden waren oder im Verlaufe der Differenzierung aktiviert wurden, kontrolliert. Somit ähnelt die direkte Programmierung von Stammzellen in Motorneuronen in vivo Entwicklungsprozessen in ihren komplexen genregulatorischen Netzwerken und ihrer Transkriptionsfaktor-Kooperativität.

Diese Arbeit zeigt wie wichtig Chromatin-Dynamik und ihre Beziehung zur Logik von Transkriptionsfaktoren ist, um die Veränderungen der Genexpression zu verstehen.

Publications

In addition to other unpublished work, this thesis includes work from the following publications. Only results directly produced by the author of this thesis are included. Contribution statements are provided in each results chapter (Chapters 4, 5, 6 and 7).

- S Velasco*, **MM Ibrahim***, A Kakumanu*, G Garipler, B Aydin, MA Al-Sayegh, A Hirsekorn, F Abdul-Rahman, R Satija, U Ohler, S Mahony, EO Mazzoni. **2016**. A Multi-step Transcriptional and Chromatin State Cascade Underlies Motor Neuron Programming. *Cell Stem Cell*, In press. doi: dx.doi.org/10.1016/j.stem.2016.11.006.

- SA Lacadie, **MM Ibrahim**, S Gokhale, U Ohler. **2016**. Divergent Transcription and Epigenetic Directionality of Human Promoters. *The FEBS Journal*, doi: 10.1111/febs.13747. [Review Article]

- SHC Duttke*, SA Lacadie*, **MM Ibrahim**, CK Glass, DL Corcoran, C Benner, S Heinz, JT Kadonaga, U Ohler. **2015**. Human Promoters Are Intrinsically Directional. *Molecular Cell*, 57(4).

- **MM Ibrahim**, SA Lacadie, U Ohler. **2015**. JAMM: a peak finder for joint analysis of NGS replicates. *Bioinformatics*, 31(1).

* = equal contribution

Erklärung

Hiermit erkläre ich, die Dissertation selbstständig und nur unter Verwendung der angegebenen Hilfen und Hilfsmittel angefertigt zu haben. Ich habe mich anderwärts nicht um einen Doktorgrad beworben und besitze keinen entsprechenden Doktorgrad.

Ich erkläre, dass ich die Dissertation oder Teile davon nicht bereits bei einer anderen wissenschaftlichen Einrichtung eingereicht habe und dass sie dort weder angenommen noch abgelehnt wurde. Ich erkläre die Kenntnisnahme der dem Verfahren zugrunde liegenden Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät I der Humboldt-Universität zu Berlin vom 27. Juni 2012. Weiterhin erkläre ich, dass keine Zusammenarbeit mit gewerblichen Promotionsberaterinnen/Promotionsberatern stattgefunden hat und dass die Grundsätze der Humboldt-Universität zu Berlin zur Sicherung guter wissenschaftlicher Praxis eingehalten wurden.

Mahmoud Ibrahim

Berlin, 8 Dezember 2016

Acknowledgments

All the work and results I report here would not have been possible without the tremendous support, help and guidance from Scott Lacadie, Silvia Velasco and especially my advisors Esteban Mazzoni and Uwe Ohler. Many others contributed to my research work directly and indirectly and I would like to thank each one of the coauthors of the research papers I contributed to over the past four years.

Ameirah made my life away from home a happy one. Thanks for that! Home is where the heart is. My family and friends back in Egypt were great and understanding and made my short trips back to Egypt a blast.

Surviving German bureaucracy and American visa waiting times is no easy task. With that I had help from Jennifer Stewart, Michaela Liemen, Sylvia Sibilak and Grietje Krabbe. Thank you.

Finally, I am very grateful to those who were kind enough to help proofread this thesis: Henriette Miko, Scott Lacadie, Hans-Hermann Wessels and Martin Burkert.

Throughout my PhD, I was funded by the Max-Delbrück-Center / New York University exchange program.

Chapter 1

Introduction

Interactions between the vast number of molecules inside the cell result in the large variation of cell types that we observe in multicellular organisms. During embryonic development, cells gradually change their morphology and function until they reach a final cell fate, such as a specific type of neuron. This process is called "cell differentiation" and involves a highly complex network of dynamic interactions between the molecules inside the cell. Understanding these networks and how they control cell differentiation will eventually allow for engineering cell differentiation processes so that we can produce specific cell types at will for drug testing and other medical purposes.

The last two decades saw huge advances in assays that enable the study of such molecular networks in a high-throughput manner. However, such data often present significant challenges in terms of processing and analysis due to their high volume, heterogeneity and noise. In this thesis, I will develop computational models for the analysis and integration of high-throughput molecular biology data. Using these computational models, I will attempt to study an engineered neuron *in vitro* differentiation system to understand the molecular control mechanisms that underlie this differentiation process.

This chapter introduces the motivation of this research work in more specific terms.

1.1 Chromatin and Gene Regulation

Cell function and response to environmental cues involve complex interactions between many proteins, RNA molecules and DNA. Genes are parts of chromosomes whose distinct order of DNA nucleotides determines which RNA molecule is transcribed from DNA, and in turn which protein is translated from this RNA molecule if any. Changes in the levels of different protein molecules can be traced in large part to changes in transcription regulation [1]. Transcription regulation refers to events that bring about changes in gene expression. Gene expression refers to how much RNA a certain gene produces relative to other genes. The first parts of genes, where transcription initiation by RNA polymerase occurs, are called gene promoters [2] and are

extremely important transcription regulatory elements. The genome also contains regions that are not necessarily used to produce functional RNA molecules but are still important for transcription regulation, usually called "distal" (ie. away from gene promoters) regulatory elements.

Genome regulatory elements are activated or repressed when they are bound by a class of proteins called transcription factors, which recognize distinct relatively short DNA nucleotide sequences, called DNA sequence "motifs", located inside the regulatory elements [3]. Transcription factors then recruit other enzymes and other transcription factors eventually leading to either recruitment of RNA polymerase to a target gene promoter and the release of polymerase from the promoter along the rest of gene to transcribe RNA (activation), or leading to stopping or inhibiting RNA transcription (repression).

Transcription factor binding and its effect on gene regulation is not determined by DNA sequence motifs alone. In eukaryotes, the genome is packaged in the nucleus together with a class of nuclear proteins called histones [4], forming "chromatin". Histones organize in octamers around which approximately 147 DNA nucleotide pairs are wound, forming "nucleosomes": the basic structural unit of chromatin [5, 6]. Nucleosomes confer structure to the genome by regulating access of transcription factors and RNA polymerase to the various genomic regulatory elements (see [7] for an example). Additionally, "histones project unstructured tails into the nuclear space. Histone tails are subject to covalent modifications, such as methylation, acetylation, and ubiquitination, which are added and removed dynamically via the help of a large group of enzymes capable of targeting specific histone tail residues with specific modifications" [8, 9].

Enhancers (a type of distal regulatory elements that is thought to enhance transcription levels from their target promoters) have been known for a long time [10–13]. Also, the relationship of histone modifications to transcription regulation has long been appreciated (see [14–16] for examples). However, the extent of the role of distal regulatory elements and their effect on gene promoter activation was realized with the advent of genome-wide high-throughput data of nucleosome positions and histone tail covalent modifications in multiple cell types, which revealed extensive regulation of chromatin structure and histone covalent modifications in both proximal and distal regions [17]. Therefore, changes in transcription regulation are the result of a complex interaction network involving DNA, transcription factors, histone modifications and other proteins like nucleosome remodeling complexes and RNA polymerase.

1.2 Transcription Regulatory Networks

A central quest in molecular biology is to understand the structure of such transcription regulation networks and to link them to changes in cell behavior. Due to technology limitations regarding the type of gene regulatory events that could be assayed, transcription regulatory networks have historically been inferred from the levels of gene

RNA molecules ignoring all the chromatin-level events and ignoring enhancers. Recently, advances in technology enabled high-throughput assays of chromatin structure, transcription factor binding and histone modifications. In this thesis, I will study cell differentiation on the molecular level and how it relates to chromatin regulation. I will do so by developing computational models that aim to analyze and integrate data obtained from those high-throughput assays. The future view is to move toward gene regulatory networks where enhancers and chromatin regulatory events are directly represented. A better understanding of these networks will eventually allow for precise engineering of differentiation systems.

1.3 Thesis Outline

The following chapter (Chapter 2) provides essential technical background both on the computational algorithmic side and on the molecular biology technology side. Chapter 3 provides an overview of the current state-of-the-art knowledge on transcription and chromatin regulation.

Chapters 4, 5, 6 and 7 provide the results of this research work. In particular, I will attempt to (1) provide a method, based on local mixture model clustering, that can accurately demarcate the widths of locations of Protein-DNA binding sites and locations of accessible chromatin from ChIP-Seq, DNase-Seq and ATAC-Seq data (Chapter 4), (2) provide a pipeline based on Hidden Markov Models to integrate multiple histone modification data sets (Chapters 4 and 5) and (3) provide a Bayesian Network method for clustering of combinatorial time-course genome-wide data such as multiple histone modification time-course data (Chapters 6 and 7).

Using those models, I will attempt to provide answers to questions regarding promoter chromatin structure, how it relates to transcription initiation patterns (Chapter 5) and how it evolves over time during cell differentiation (Chapter 6), as well as the relationship of enhancer chromatin dynamics to promoter chromatin dynamics during differentiation and how it relates to transcription factor binding (Chapter 7).

Finally, Chapters 8 and 9 provide a brief discussion and summary of the results of this work.

Chapter 2

Technical Background

2.1 Introduction

This chapter provides essential background and introductory material that serves to put the general topic of this thesis (explained in Chapter 1) in a more technical context. The chapter starts with a short primer on Bayesian Networks, mixture models and Hidden Markov Models, followed by an introduction to high-throughput genome-wide sequencing techniques that profile gene expression, DNA-protein interactions and chromatin structure and finally a discussion on important concepts pertaining to genome-wide data analysis.

2.2 Mixture Models and Bayesian Networks

2.2.1 Model-based Analysis

In model-based analysis, one assumes that the observed data is produced from an underlying probabilistic model. In the simplest case scenario, this probabilistic model is a univariate probability distribution, such as a univariate Gaussian (normal) distribution $\mathcal{N}(\mu,\sigma)$, where μ is the mean of the distribution and σ is the standard deviation of the distribution. Probability distributions describe the likelihood of observing a certain outcome using the Probability Density Function (PDF). PDFs which are defined as a function of the probability distribution parameters (like μ and σ in the case of the Gaussian distribution). For example, the PDF of a Gaussian distribution $\mathcal{N}(\mu,\sigma)$ is defined by

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$f(x|\mu,\sigma)$ gives the relative likelihood of observing a continuous number x given a Gaussian distribution $\mathcal{N}(\mu,\sigma)$.

It is often the case that some number n of observations $x_1 \ldots x_n$ had been collected, and we assume that such data can be modeled using a certain PDF, say a Gaussian

PDF. Since the *true* μ and σ are not known, one wants to *estimate* them. The task is then to estimate the distribution parameters μ and σ that produce a PDF that best explains the observations $x_1 \ldots x_n$. A standard method to estimate the parameters is the maximum-likelihood method (ML) which maximizes the likelihood of the values of the parameters μ and σ given the data $x_1 \ldots x_n$. In other words, we want to derive a function that gives the likelihood of the model given the data, and we then should find the value(s) of the probability distribution parameters that maximize this function. To avoid underflow issues, this is usually done in the log space. In the case of the Gaussian distribution, the log-likelihood function $\mathcal{L}$ is defined by:

$$\mathcal{L}(\mu, \sigma | x_{1\ldots n}) = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

The Gaussian distribution is especially convenient because there exists a simple *analytical solution* to maximize the values of μ and σ that maximize this likelihood function (this is often not the case when working with other distributions). The ML estimates of μ and σ are given by $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$, which are the sample mean and the sample variance.

2.2.2 Mixture Models

In many cases, one assumes the collected data $x_1 \ldots x_n$ arises from two or more different processes and therefore can only be modeled as a *mixture* of two or more (Gaussian) distributions $\mathcal{N}_1$ and $\mathcal{N}_2$, each with a different set of model parameters. The Gaussian distributions in the mixture are often called "mixture components". However, the parameters defining the two distributions $\mathcal{N}_1$ and $\mathcal{N}_2$ do not uniquely define a mixture model of the two distributions. A mixture of Gaussians additionally requires a vector of *mixture weights* $w_1 \ldots w_g$ where g is the number of mixture components in the model. $w_1 \ldots w_g$ define the relative amount of data each distribution contributes to the mixture and it must sum to 1 over all g. More formally, it defines the *prior* probability (meaning the probability that we would have assigned to any data point if we have not observed it yet) that any data point from the mixture belongs to one of the distributions in the mixture $\mathcal{N}_j (j = 1, \ldots, g)$.

Therefore, a Gaussian mixture model $\mathcal{M}$ with g mixture components is fully defined by the parameter set:

$$\mathcal{M} \sim \left(w_{1\ldots g}, \mu_{1\ldots g}, \sigma_{1\ldots g} \right)$$

and the PDF of a Gaussian mixture model is defined by:

$$f(x|\mathcal{M}) = \sum_{j=1}^{g} w_j f(x|\mu_j, \sigma_j)$$

where $f(x|\mu_j, \sigma_j)$ is the likelihood of observing a continuous number x given the mixture component $\mathcal{N}_j$.

2.2.3 The Expectation-Maximization Algorithm

Estimating the parameters of mixture models is more complicated than that of one univariate model. It is usually not possible to do so analytically, even for mixtures of Gaussians. The ML estimates of mixture model parameters are often obtained using a well-known iterative algorithm called the Expectation-Maximization (EM) algorithm [18]. With the EM algorithm for mixture models, one typically assumes that there is an unobserved (ie. *hidden*) class variable $\mathcal{Z}$ upon which the observed data $x_{1\ldots n}$ is conditioned. One assumes a certain number of values g that $\mathcal{Z}$ can take, which is the same g that is the number of mixture components we assume the data arises from. $\mathcal{Z}$ can take each value j with a certain probability and those probabilities sum to 1. Therefore, $\mathcal{Z}$ defines the mixture weights $w_{1\ldots g}$, which is the same as the *prior* probability of observing a data point from a mixture component $\mathcal{N}_j$. Note that the unobserved variable $\mathcal{Z}$ was added to our model definition but without introducing any new model parameters, it simply adds "context" to the mixture weights and allows us to reformulate the mixture model as one containing both observed and unobserved variables. The EM algorithm is designed specifically for getting the maximum likelihood estimates of models containing both observed and unobserved variables.

The EM algorithm starts from a certain initial estimate of the model parameters. We denote the set of model parameter estimates as θ and this particular initiation parameter set the algorithm starts from as θ_{init}. The EM algorithm then proceeds to update θ iteratively by going through the expectation (E-step) and maximization (M-step) steps. In the E-step, the likelihood of each data point x_i being produced from each mixture component j is calculated. In the M-step, the model parameters are updated given the likelihood estimates from the previous E-step.

E-step

In the case of a Gaussian mixture model, the E-step can be expressed as calculating each "membership value" v_{ij} for each data point x_i and mixture component j (ie. each possible value of $\mathcal{Z}$). The membership value is the probability that the unobserved variable Z takes some value j given the value of some data point x_i and some parameter set values θ

$$v_{ij} = p(Z = j|x_i, \theta)$$

Note that in the case of a Gaussian mixture θ denotes the mixture weights and the means and variances of the Gaussian mixture components (see above):

$$p(Z = j|x_i, \theta) = \frac{w_j f(x_i|\mu_j, \sigma_j)}{\sum_{l=1}^{g} w_l f(x_i|\mu_l, \sigma_l)}$$

The denominator is a normalization factor that ensures that for each data point x_i, the sum of v_i across all j values of Z is equal to 1: $\sum_{j=1}^{g} v_{ij} = 1$.

M-step

In the M-step, we update the parameter set θ according to the new membership values v_{ig}. If $m = i \cdot g$, θ is updated by:

$$w_j^{new} = \frac{\sum_{i=1}^{n} v_{ij}}{\sum_{j=1}^{g}\sum_{i=1}^{n} v_{ij}}$$

$$\mu_j^{new} = \frac{\sum_{i=1}^{n} v_{ij}.x_i}{\sum_{i=1}^{n} v_{ij}}$$

$$\sigma_j^{2^{new}} = \frac{\sum_{i=1}^{n} v_{ij}(x_i - \mu_j^{new})^2}{\frac{m-1}{m}.\sum_{i=1}^{i} v_{ij}}$$

One iterates through those two steps until there is no appreciable change to the log-likelihood of the model parameters given the data: $\mathcal{L}(\theta|x_{1\ldots n})$:

$$\mathcal{L}(\theta|x_{1\ldots n}) = \sum_{i=1}^{n}\left(log\sum_{j=1}^{g} w_j f(x_i|Z=j,\theta)\right)$$

The EM algorithm might converge to a local maximum, meaning that the values of the parameters found by the EM algorithm are not necessarily the absolute best fit of the model to the data. The parameters determined by the EM algorithm will depend on the values of θ_{init}: the initial parameter estimates that were chosen to start the first E-step. Therefore, it is usually advised to run the EM algorithm multiple times from multiple initializations or to determine a meaningful method for initializing it. In the case of Gaussian mixtures, θ_{init} is often determined based on k-means or hierarchical clustering of the data.

Estimating the parameters of a mixture model using the EM algorithm is often called "model-based clustering" [19] because for each data point it assigns a probability of belonging to one of the mixture components (ie. one of the clusters or classes) and this probability is based on the assumption of an underlying probabilistic model (ie. model-based). Throughout this thesis, we will employ model-based clustering via the EM algorithm, and variants of it, to learn the parameters of mixture models (Chapter 4) and other more complex models (Chapters 4, 6 and 7).

2.2.4 Bayesian Networks

Mixture models can also be depicted in the form of Bayesian Networks (BN) [20] which are directed acyclic graphs (DAGs) that define the relationships between random variables. In a BN, nodes represent random variables and edges represent conditional dependencies between those variables. Thus, in the example BN in Figure 2.1a, $\mathcal{N}_i$ and $\mathcal{N}_j$ are independent of each other if the value of C is known. This conditional independence assumption is why this particular structure is called a *Naive* Bayes model. The Gaussian mixture model described above can also be depicted as a BN (Figure 2.1b): the values of the Gaussian parameters μ and σ are determined by the value of

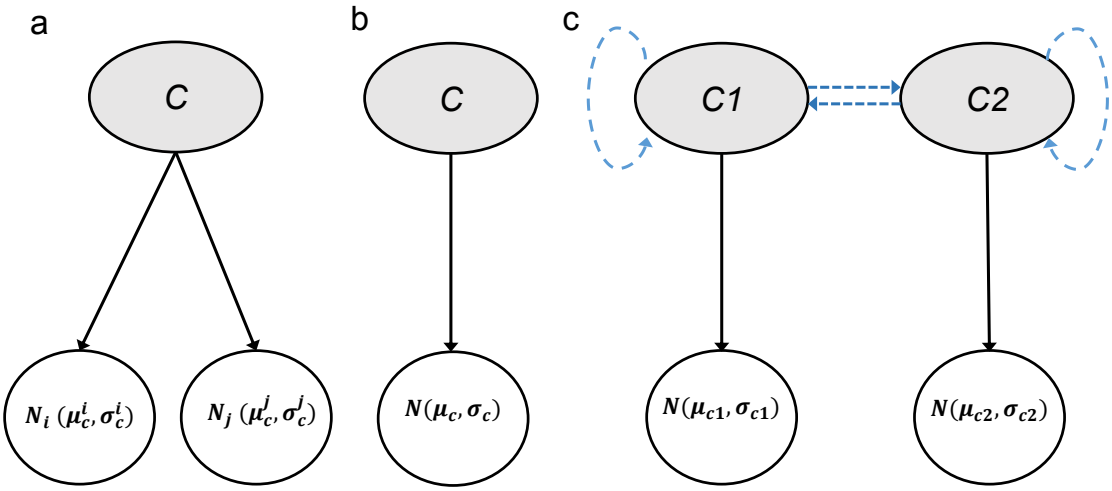


Figure 2.1: Bayesian Network representations of (a) a Naive Bayes model with two Gaussian features and (b) a Gaussian mixture model. (c) shows an unrolled Hidden Markov Model with two states and univariate Gaussian emissions.

the class variable C. In other words, there are different values of μ and σ for each value of C. The difference between a naive Bayes model (Figure 2.1a) and a mixture model (Figure 2.1b) is that the mixture model defines one random variable that is generated from a mixture of multiple different distributions (processes), while the naive Bayes model defines multiple random variables whose distribution parameters are depedent on the class (value of C) but are otherwise independent of each other (*conditionally independent*).

In BNs we say that a variable x is a parent of a variable y and y is a child of x if there is a directed edge going from x to y. For example, in Figure 2.1a the variable C is a parent of N_i and N_j. A very useful understanding of how BNs encode a PDF of a model is to understand factorization via the chain rule: if $\mathcal{B}$ is a BN with n random variables $x_{1\ldots n}$ then the PDF of $\mathcal{B}$ can be written as:

$$f(\mathcal{B}) = \prod_{i=1}^{n} p\left(x_i | x_{pa(i)}\right)$$

where $x_{pa(i)}$ are the parent variables of x_i. Therefore, expressing a statistical model as a BN allows for factorizing the joint probability space of the model in terms of the "local" conditional probability constraints between the variables. In Chapters 6 and 7, we will design a BN model to cluster and integrate time-course highly-dimensional genome-wide data.

2.2.5 Hidden Markov Models

A well-known class of (dynamic) BN models that has become extremely popular in many fields including Bioinformatics is Hidden Markov Models (HMM). HMMs describe what essentially is a mixture model but with additional restrictions. It often helps to explain HMMs as a model for temporal processes. The main idea is that the likelihood of the unobserved class variable Z taking a value k at time-point $t+1$ is conditioned upon its likelihood at time-point t. Therefore, HMMs are mixture models with an extra parameter set, called the *transition probability matrix*, which define the probability of a certain value of Z at time $t+1$ given its value at time t. Figure 2.1c shows an example of an unrolled HMM (by "unrolled" we mean that the values the class variable C can take are explicitly depicted). This HMM models one Gaussian random variable generated from two classes (called "states" in the context of HMMs). The dashed blue edges represent the transition probabilities. Note that the transition probabilities do not replace the *prior* probabilities (what we called mixture weights above). The prior probabilities are called "initial probabilities" in the context of HMMs and define the probability of observing C taking a certain value before the process had started. Therefore, an HMM model $\mathcal{H}$ with n states $s_{1\ldots n}$ is defined by the following parameter set:

$$\mathcal{H} \sim \left(e_{1\ldots n}, i_{1\ldots n}, A = [a_{jk}]\right)$$

where n is the number of states, i are the initial state probabilities where $\sum_{n=1}^{n} i_n = 1$, e are the observation probability distribution parameters (called "emission distribution", μ and σ in the case of Gaussian emissions) and A is the transition probability matrix. A is a square matrix where each element $a_{jk} = p(q_{t+1} = s_k | q_t = s_j)$ where $1 \leq j,k \leq n$ and q_t is the state at time t.

The parameters of HMMs that maximize the likelihood of the model given the data can be found with a variant of the EM algorithm called the Baum-Welch algorithm, explained elegantly in the well-known tutorial on HMMs by Lawrence R. Rabiner [21]. The Baum-Welch algorithm utilizes a dynamic programming algorithm called the Forward-Backward algorithm to obtain the posterior probabilities of all possible hidden state assignments given a sequence of observations.

Another important task with HMMs is to infer the most likely state path over time given the data and an HMM model parameter set $\mathcal{H}$. This is usually called "decoding" and is accomplished using the Viterbi algorithm [22]. Note that unlike the Forward-Backward algorithm, the Viterbi algorithm operates on a "left-to-right" fashion computing the *most likely* state path. When the Forward-Backward algorithm is used for decoding instead, it is often called "posterior decoding".

In genomics, HMMs are often used to represent sequence along the DNA rather than time. For example, in the problem of computationally defining gene locations in the genome, the gene components (UTRs, exons, introns) and their expected order can be modeled using HMMs [23]. In Chapter 4, we will consider HMM models that integrate multiple genome-wide data to summarize chromatin and transcription information along the genome length.

2.3 Genome-wide Sequencing Assays

The past decade saw an increasing number of methods that rely on high-throughput sequencing data to profile a certain aspect of cell function in a high-throughput genome-wide manner. In fact, sequencing-based assays are now ubiquitous with maybe hundreds of variants measuring many different aspects of cell biology. The general idea is to purify a population of hundreds of millions of DNA molecules, which represent a certain property of interest (for example those bound to a protein of interest or those resulting from reverse transcription of certain RNA molecules). Those DNA molecules are then sequenced using massively-parallel sequencing platforms. The obtained sequences are usually called "reads". When reads are aligned back to the genome, the number of reads that align to a certain genomic location can indicate a quantitative assessment of a transcription factor binding to that location or how much RNA that location produces...etc.

Because of how most genome-wide sequencing assays are designed, one can think of sequencing reads as being collected *randomly* from the all genomic locations and from all assayed cells, but in a manner that is biased such that a genomic location of interest (bound to a certain transcription factor for example) is more likely to generate a higher proportion of the reads than other locations (those not bound to the assayed transcription factor). The number of reads aligned to a certain location can therefore be thought of as a *probabilistic* measure of the property being assayed at a given genomic location. In most, if not all, sequencing assays this probabilistic measure depends on many factors that are difficult or even impossible to control, making high-throughput sequencing assays inexact or "noisy" at best.

We will focus below on assays related to chromatin structure and transcription quantification (following sections), paying particular attention to the possible sources of noise and uncertainty in such data. But first we enumerate sources of uncertainty common to all sequencing assays produced in cell populations:

1. Excluding recent single-cell assays, all high-throughput sequencing assays are performed on a population of hundreds of thousands of cells (ATAC-Seq) to up to 50 million cells (DNase-Seq). Although one strives to ensure that all cells assayed are homogeneous and are in the same state, there is no guarantee that on the molecular level all cells are exactly the same. For example, a certain genomic location might be bound by a transcription factor in only a fraction of the cells. The cell population is also one of the main reasons one observes a continuous distribution of ChIP-/DNase-Seq signal intensities rather than a binary bound/not bound result.

2. In many cases, the genome is assumed to be of certain ploidy. However, this assumption is often not true in *in vitro* cell culture systems and especially in cancer models where different regions in the genome might have different duplications or deletions, often referred to as copy number variations [24]. copy number variations are often estimated from data if they are not directly known and accounted for using various models (see [25] for an example).

3. Before sequencing, DNA molecules need to be attached to amplification primers and then amplified using PCR in order to ensure that there is a sufficient amount of DNA to be detected by sequencing. Three sources of biases are introduced: the amplification primers might prefer certain DNA molecules relative to others, amplification via PCR introduces a GC bias where DNA molecules rich in GC content are more preferably amplified [26] and finally amplification might lead to PCR-duplicates where the same exact DNA molecule is sequenced multiple times. When one observes two reads that are on the same strand and align to exactly the same location, it is impossible to tell whether they originate from two different original DNA molecules or from PCR-duplicates. Unique molecular identifiers are randomized amplification primers that can be used to remedy some PCR amplification issues [27]. Otherwise, PCR-free protocols have also been developed.

4. Sequencing depth is perhaps the most important factor to consider in high-throughput sequencing assays. Sequencing depth is how many basepairs of the genome were covered by the DNA reads sequenced, on average. For example, in ChIP-Seq it is calculated as $(n * f)/g$, where n is the number of reads sequenced, f is the average fragment length and g is the genome size. The more reads sequenced, the higher the depth and the higher the likelihood of capturing all binding sites / open regions / RNA molecules..etc. to their correct relative abundances. One can also think of this in terms of sequencing experiments being a random sampling of the property being assayed. The higher the sampling rate, the more likely one is able to recover the *true* probability distribution. It is worth noting that for organisms with large genomes like human and mouse, it is often impossible to sequence the sample to enough depth to guarantee that the read sampling rate was sufficient. The effect of sequencing depth variability on ChIP-Seq is extensively explored in [28].

5. Finally, it should be noted that sequencing technology is not perfect. When a DNA molecule is sequenced, errors can occur in the base calls, meaning that there can be errors in the output sequence. Depending on the sequencing technology, various base quality scores can be computed and used to determine whether overall the read sequence determined is of acceptable quality. Generally, this issues leads to uncertainty in determining the genomic location the read comes from with absolute certainty.

2.3.1 Assays of Chromatin Structure

ChIP-Seq [29, 30] (Protein-DNA Binding, Fig. 2): This assay requires an antibody that is specific to a certain protein of interest like a transcription factor or a histone with a certain post-translational modification. The first step is to cross-link all DNA-bound proteins to the DNA thereby fixing those interactions, akin to freezing those interactions in time. This is usually accomplished using formaldehyde treatment. The

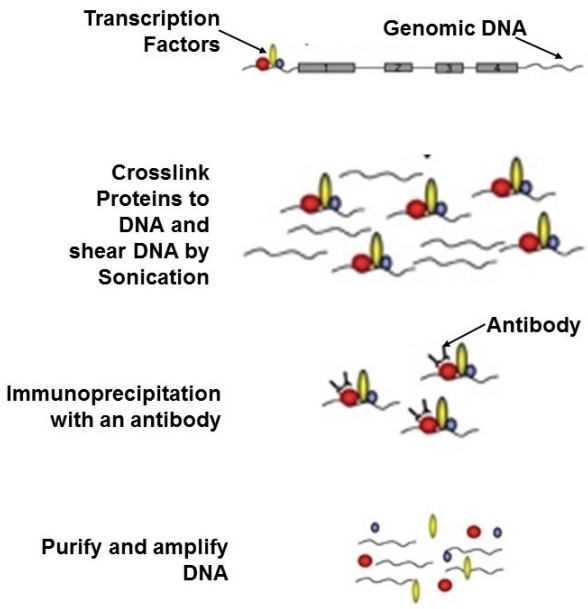


Figure 2.2: An overview of ChIP-Seq Protocol. Figure adapted from Tam, W.-L. and Lim, B., Genome-wide transcription factor localization and function in stem cells (September 15, 2008), StemBook, ed. The Stem Cell Research Community, StemBook, doi/10.3824/stembook.1.19.1, http://www.stembook.org.

next step is to randomly fragment those DNA-Protein complexes via sonication or enzymatic digestion. The DNA molecules resulting from the fragmentation process are usually called "input" and often used as background or a "null" assay. The fragments of the Protein-DNA complexes that contain the protein of interest are isolated using antibodies specific to that protein and purified. The proteins are reverse-crosslinked from DNA and the remaining DNA fragments are sequenced. When those sequences of DNA fragments are aligned back to a reference genome, the locations in the genome to which the protein of interest was bound can be identified. Prominent recent advances in ChIP-Seq technology include higher resolution variants (ChIP-exo [31] and X-ChIP [32]), single-cell ChIP-Seq [33] and a technology for assaying for two proteins co-bound to the same locations called Co-ChIP [34].

Sources of uncertainty in ChIP-Seq data:

1. It is important to note that since genomic locations bound to large protein complexes are protected from sonication and enzymatic digestion. The ends of ChIP-Seq "input" fragments might be enriched in locations not bound by large protein complexes [35]. Therefore, input fragments are not guaranteed to be uniformly distributed across the genome but might favor open chromatin thereby biasing ChIP-Seq against genomic domains that are "closed". This is perhaps one contributing factor to the observation that locations with tightly packed arrays of nucleosomes result in broad low signal-to-noise ratio signal when assayed with ChIP-Seq, rather than sharp signal. Furthermore, sonication is affected by the 3D

structure of the genome in ways that are not yet understood. Finally, sonication time and detailed parameters affect the extent to which the DNA is fragmented. Although most researchers seem to isolate fragments with a certain length range after sonication to avoid such variability, differences in sonication strategies can still introduce variabilities and biases that are not understood.

2. After fragmentation, an antibody is used to select DNA fragments bound to a protein of interest. This introduces an important source of uncertainty since the specificity of antibodies is variable between different antibodies with some having more off-targets than others [36]. Furthermore, polyclonal antibodies (typically used in ChIP-Seq) vary between lots since each lot is obtained from a different animal [36]. Monoclonal antibodies are obtained from a single purified cell line *in vitro* and can overcome polyclonal antibody limitations to a great extent [37]. [38] provide a database of commercially available antibodies and their specificities. One should still note however, that an antibody can recognize a target protein in a specific conformation but not in another. So for example, an antibody can fail to recognize a certain histone acetylation if there is also methylation on the same histone or fail to recognize a transcription factor when it is cobound with another factor. Therefore, all such aspects introduce problems in the interpretation of the data. Ideally, researchers would reproduce their data using different antibodies to ensure that the result is not an artefact of the antibody used.

3. The antibody-bound DNA molecules need to be purified. This purification step often involves magnetic beads. The primary antibody is incubated with magnetic beads such that the antibody-bound chromatin fragments can be attached to those beads as well and then isolated using what essentially is magnetic chromatography: the bound chromatin is stuck to the wall of the test tube using a magnet and unbound chromatin is eluted. This step is repeated 2 or 3 times. However, it is of course not perfect. DNA molecules that were highly abundant in input chromatin might remain in the final isolated fraction even if they not bound by the antibody, leading to "phantom" ChIP-Seq peaks [39] and DNA molecules that were antibody-bound might get erroneously completely eluted if they were lowly abundant in chromatin input.

4. ChIP-Seq DNA fragments are typically around 150 basepairs to 300 basepairs in length. This dictates the resolution of ChIP-Seq. When data from multiple fragments are aggregated, the resolution of ChIP-Seq can be improved to approximately 50 basepairs in aggregate. But fragment lengths present another problem: since sequencing reads are typically short, only the starts of the isolated fragments are sequenced, giving rise to a pattern where the location of the bound factor is depleted from reads but the locations adjacent to it are enriched in plus-stranded reads on one side and minus-stranded reads on the other side. In the analysis of ChIP-Seq data, an average fragment length is typically estimated and the reads are extended to that length or shifted by half of that length. Of

course, this average fragment length is an approximation since the actual fragments have a probabilistic unknown distribution. This problem can be resolved using paired-end sequencing where the length of each fragment can be measured.

DNase-Seq [40, 41] and ATAC-Seq [42] (Chromatin Structure): This is in fact another variant of protein-DNA binding assays, except that the focus here is on genomic locations where proteins are *not* bound. Those assays usually capture genomic locations that are depleted of nucleosomes and other protein complexes of comparable sizes. This is often extremely useful because such genomic locations are usually active regulatory elements regulating gene expression such as enhancers and promoters. In DNase-Seq, the DNA is digested using the DNaseI enzyme which will preferentially cut the DNA in locations where the DNA is "accessible" or "open" (that is, not bound by a protein or other molecules). The ends of the resulting fragments are sequenced and mapped back to a reference genome thereby providing a measure of genome "accessibility" on a local level. In ATAC-Seq, the same approach is followed except that a hyperactive version of the Transposase Tn5 is used to fragment the DNA instead of DNaseI. The advantage of ATAC-Seq is that it requires far fewer cells than DNase-Seq and is far less tedious. Other assays for DNA accessibility include FAIRE-Seq [43] and Sono-Seq [35]. It should also be noted that although those assays intend to measure the same property, at least conceptually, they in fact result in different profiles of "accessibility" due to the vast differences between the protocols.

Sources of uncertainty in DNase-/ATAC-Seq data include:

1. DNase-Seq and ATAC-Seq are essentially enzymatic digestions of chromatin, and they also are potentially affected in an unknown manner by the 3D structure of the genome like sonication in ChIP-Seq.

2. Also like sonication, the concentration of the enzyme and the incubation time with chromatin will affect the amount of digestion and the fragment length distribution obtained. In DNase-Seq one selects a perceived optimal digestion pattern from a pulsed field gel [44]. In ATAC-Seq, one manipulates the Tn5 concentration, the number of cells used and the incubation time to obtain a desired fragment length distribution. However, this is often somewhat subjective.

3. DNaseI and Tn5 transposase are not completely random in their targeting of DNA but have specific sequence preferences [45, 46]. Meaning that they prefer to recognize and cut certain DNA sequences over others. Therefore, this biases the overall distribution of fragments obtained from those assays toward open regions containing the preferred sequences [46]. It is important to note that such sequence preferences might also be dependent on cell type and experimental condition [47]. Recently, researchers have attempted various strategies to correct for DNase-Seq and ATAC-Seq sequence bias, especially in the context of transcription factor footprinting (see [48, 49] for examples).

4. It is important to note that protein DNA contacts are not fixed. Transcription factors are continuously binding and leaving their binding site with different factors

having different residence times [50]. It is possible to envision that factors with higher residence time and locations where a factor has higher residence time have a higher contribution to ChIP-Seq signal than other locations. This is also relevant for histone modification ChIP and for DNase-Seq; a location with stable nucleosomes that do not change their locations can contribute higher clearer ChIP/DNase signal than another location where nucleosomes are unstable or are more frequently remodeled.

2.3.2 Assays of Transcription

GRO-Seq [51] (Nascent Transcription): The goal here is to assay nascent RNA molecules "just transcribed" by polymerase. The nuclei are isolated and incubated with BrdUTP and Sakrosyl which inhibits attachment of new polymerase molecules. Only currently engaged polymerase molecules produce BrdUTP labeled RNA molecules which can then be isolated using an anti-BrdUTP antibody when PolII is released. Those RNA molecules are reverse transcribed to DNA and sequenced. Mapping those reads back to a reference genome provides a snapshot of the nascent RNA production that was occurring in the nuclei when they were isolated. A high-resolution variant of GRO-Seq is PRO-Seq [52]. It should be noted that this is essentially an *in vitro* transcription assay and transcription occurring in such artificial conditions might cause other unanticipated changes in transcription regulation. Alternatives include NET-Seq [53], Nascent-Seq [54] and TT-Seq [55] which relies on 4su labeling of transcribed RNA.

GRO-Cap [56] (Nascent Transcription Initiation): This is the same as GRO-Seq except that it contains an extra reaction to enrich for 5'CAP of RNA molecules, thereby assaying transcription start sites of nascent RNAs.

RNA-Seq [57] (Transcript Abundance): In this assay, RNA molecules are isolated, reverse transcribed and sequenced. This is perhaps the most common high-throughput sequencing assay and the most standardized in terms of bench-side protocols and analysis methods. Many variants exist. For example, RNA in a certain fraction of the cell (chromatin, nucleus, cytoplasm..etc.) can be isolated, RNA with polyA tails can be isolated, short RNA molecules can be enriched for...etc. Although often interpreted as an assay of transcription, it should be noted that unless one enriches specifically for nascently transcribed RNA (see above), RNA-Seq represents the steady-state population of RNA molecules found in the cells and is the result of RNA production, processing and degradation processes. Therefore, RNA-Seq of steady-state RNA species is a mid-way view of gene regulation that is downstream of transcription regulation.

2.4 Analysis of Genome-wide Sequencing Assays

2.4.1 Read Alignment

The first step in the analysis of high-throughput sequencing assays is to align the reads back to a reference genome. Read alignment in its simplest form involves two steps: (1) generating a dictionary of all genomic locations and their sequences, often called a genome index and (2) for every read, the genome index is searched to find a match so that the read is assigned to that location. Because of base call errors and deviations from the reference genome, one would not expect to find exclusively exact matches but it is often the case that a read maps to a certain location with one, two or more base mismatches. Furthermore, because sequencing reads are short (in the order of tens of basepairs) an issue arises where a read can match multiple locations in the genome, since a string of say 50 basepairs with a certain sequence might occur multiple times in the reference genome index. This issue is often called "mappability" and it refers to whether sequencing reads can be mapped uniquely to a certain location of the genome given a certain read length. Genomic locations with repetitive sequences have low mappability, so it is difficult to assign reads to such locations with certainty. Read aligners can combine mismatch information and the number of locations a read aligns to in order to calculate a read alignment score, which can then be used to filter for reads that were confidently aligned. Alternatively, it is common to also directly remove reads that have more than a certain number of mismatches and/or those that aligned to multiple locations. Different aligners differ in the heuristics they use to resolve read alignment uncertainties and in the algorithms they use to search the genome index. Popular programs for ChIP-/DNase-/ATAC-Seq include Bowtie [58], Bowtie2 [59] and bwa [60]. When aligning reads to transcripts as in RNA-Seq assays, one needs to consider transcript splicing: a read obtained from a mature transcript can be split in the reference genome if it covers the end of one exon and the beginning of another. Programs that can resolve this issue include Tophat [61], STAR [62] and RSEM [63] which uses bowtie to align reads to the transcriptome and quantifies expression using read counts within an Expectation-Maximization framework.

Once reads have been aligned to the reference genome, one can count how many reads aligned to a certain genomic location. This is the most common and most ubiquitous method of summarizing genome-wide sequencing data. Therefore, these data types can be considered count-type data, which can be summarized using count-type probability distributions such as the Poisson, Multinomial and Negative Binomial distribution. This data can also be made to behave in a more continuous manner in order to use the more convenient Gaussian distribution. Commonly, sequencing read counts can also be smoothed along the length of the genome in order to produce an easier to interpret ChIP-/DNase-Seq "signal" (see [64] for an example).

2.4.2 Finding Enriched Regions in Genome-wide Data

One of the most common tasks in the analysis of ChIP-/DNase-/ATAC-Seq data is the task of determining enriched locations in genome-wide sequencing data. This is usually called "peak finding" and can be restated in a more formal fashion as discretizing the genome into regions where there is a significant presence of sequencing reads (ie. peaks) and regions where there is not. This task can be approached in four different ways: (1) finding regions that are significantly different from background assuming a certain background read count probability distribution, (2) segmenting the genome into enriched and not enriched regions, (3) exploiting a certain property of a specific experimental protocol to find peaks, like for example a peak finder for transcription factor ChIP-Seq and (4) detecting regions that contain a certain signal shape using signal detection methods.

One of the most popular peak finders is MACS [65] which falls in the first category and assumes read counts follow a local Poisson distribution. MACS starts by binning the genome into overlapping bins and obtains the read counts in those bins. Then compares the Poisson rate parameter estimated from the sample counts in each bin λ_{fg} to that obtained from neighboring bins, and to that obtained from the background experiment (example: ChIP-Seq input) if available, λ_{bg}. This way MACS can assign a p-value to each bin the genome. This procedure introduces two simple but important concepts: (1) genome binning and (2) modeling read count distributions in genome-wide sequencing data. Genome binning is done because of the computational burden involved and because ChIP-Seq and similar protocols are usually not single-base resolution protocols. Scanning the genome in bins introduces the question of which bin size to use and leads to coarse identification of binding sites since the bin identified as statistically significant will include a peak but will not accurately demarcate that peak's width. Modeling read count distributions is desired if one has to derive a p-value expressing the confidence in the peak being "real/true" as opposed to resulting from noise. The most natural approach for modeling read counts is to consider a Poisson distribution with rate λ. However, read count data are over-dispersed, meaning that it is not possible to model all bins in the genome with the same λ. MACS resolves this issue by estimating a different local background λ_{bg} for each candidate peak separately. A more elegant solution is to use the negative binomial distribution or its zero-inflated version (to account for the presence of a lot of locations in the genome without any mapped reads) [66] or to use a Hidden Markov Model (HMM) with 2 or 3 hidden states [67].

Peak finders relying on HMMs usually assume that read counts arise from two or three distributions representing peaks, ambiguous regions and noise. It is generally accepted that using an HMM with only 2 states would identify ambiguous and peak regions in one state. One of the main applications of HMM peak finders is to find peaks in ChIP-Seq data with large enriched domains featuring low Signal-to-Noise Ratio (SNR) like H3K27me3 and H3K36me3 ChIP-Seq data [67], because constraining the bin assignment to the hidden state by the HMM transition probabilities helps avoid calling many fractured peaks instead of a supposedly more appropriate one large

enriched domain. A similar approach is Change Point detection which, like HMMs is rooted in time-series analysis, works especially well in defining large enriched domains [68].

An alternative to statistically-rigorous models, many programs focus on a specific data type and exploit a property of the experimental protocol to find peaks. The most common example of this are peak finders for transcription factor ChIP-Seq which rely on identifying the plus-strand/minus-strand bimodal distribution expected in ChIP-Seq data. These peak finders can identify transcription factor binding sites at very high-resolution, close to one-basepair resolution when analyzing ChIP-exo data [69–71]. The presence of punctate ChIP-Seq data types (transcription factors), broad data (large-domain histone modifications) and data that is in between like Polymerase ChIP-Seq invites the question whether one method can identify peaks reliably in all three cases using the same approach?

DFilter [72] is a peak finder that frames peak finding as a signal detection problem and is an attempt at a generally-applicable peak finder. The idea is that once the desired signal shape (read count local spatial distribution) is found, this signal shape can be reliably located in a genome-wide manner regardless of what that shape actually is. This is an elegant approach except for the fact that one needs to know beforehand which signal shape to look for. This can be trivial in some cases but not in broad data and will fail to generalize when both types of signals are found in the same data set (as would be expected in Polymerase ChIP-Seq data for example). In Chapter 4, I will introduce JAMM (Joint Analysis of NGS replicates via Mixture Model clustering): an attempt at a universal peak finder that can find accurate peak boundaries regardless of signal shape while utilizing all available biological replicates at once [73].

2.4.3 Normalization of Genome-wide Data

Analysis of genome-wide high-throughput sequencing data poses significant challenges in terms of data integration, especially if data across multiple time-points and/or multiple conditions need to be integrated. When reviewing these issues here, I will frequently refer to gene expression and genes as the entities of interest although the same concepts apply to high-throughput sequencing technologies where the entities of interests are genomic locations like promoters or ChIP-Seq peaks.

An obvious issue in analyzing multiple data sets is sequencing depth variability. A gene g whose real expression value (number of transcripts) does not change between two conditions $C1$ and $C2$, measured by two RNA-Seq samples $S1$ and $S2$ respectively, might have 10 reads in $S1$ and 100 reads in $S2$ if $S2$ has 10x the number of reads in $S1$. A straightforward correction value is depth normalization which is simply dividing the number of reads assigned to g by the total library size. This is essentially a linear transformation of the data: $Sn_i = S_i \times f_c^{-1}$, where S_i is the raw read count per gene vector, f_c^{-1} is the reciprocal of the correction factor (in depth normalization, the sum of S_i) and Sn_i is the normalized read count per gene vector.

Variations on this approach exist such as using the top 75 % quantile to calculate f_c in order to avoid influence of very lowly expressed genes [74], or using the median

of counts [75], or additionally normalizing for the gene length giving the popular measurement Reads Per Kilobase of transcript per Million mapped reads (RPKM, [57]). A strong assumption in depth normalization is that the global transcriptome repertoire and global quantity of RNA is essentially the same between conditions. This is of course not always true. Condition $C1$ could express the same genes at the same level as $C2$ but have additional $C1$-uniquely expressed genes. In this case, depth normalization will lead to falsely identifying common genes as down-regulated in $C1$ relative to $C2$ and will underestimate the extent to which $C1$-unique genes are up-regulated. Trimmed mean of M-values (TMM) [76] and DEseq [77] are normalization methods that address this particular issue. The common assumption in these two methods is that the real expression of the vast majority of genes between these two conditions will not change, hence, f_c is calculated as a summary statistic of the distribution of the gene-wise fold-change values between the two conditions. In the case of TMM, it is a weighted average, by the absolute expression value of each gene [76], and in DESeq it is the median [77].

What if we no longer believe that the vast majority of genes have the same expression value in both conditions? The answer comes with non-parametric rank-based normalization methods such as rank normalization [78] and quantile normalization [79]. In those methods, we are no longer interested in estimating a single correction factor to linearly transform the data. Rank normalization simply disposes of the actual expression values and preserves only the rank of each gene in each sample [78]. Quantile normalization matches the empirical distributions of gene expression between the different conditions (can be understood as transforming the data so that a quantile-quantile plot is diagonal) [79]. Since the quantile and rank normalization methods do not use a gene-by-gene value but rather match gene ranks, there is no assumption that the vast majority of genes have the same expression values across conditions, but there is another strong assumption that the overall gene expression distribution in both conditions is the same. In fact expression values in one sample become simply a permutation of the other. This assumption might not be always valid and, like with any other normalization method, can lead to potential issues in downstream analysis if it does not hold [75].

When applying all these normalization methods we are not interested in whether cells in condition $C2$ produce half or twice the amount of RNA cells in condition $C1$ produce. In some cases, such global effects are interesting. For example, cells with higher amounts of c-myc exhibit amplified amounts of gene expression in general [80, 81] and a comparison between those two conditions using any of the normalization methods ignores this effect [82, 83]. A catchall solution seems to be spike-in controls, which would allow for an estimation of the actual true abundance of an RNA transcript from high-throughput sequencing experiments based on interpolation to a standard curve [82, 84].

ChIP-Seq and DNase-/ATAC-Seq data as well can be normalized using spike-ins [85–87], although this is not yet widely adopted. When no spike-ins are available any of the aforementioned normalization methods can be used. However, interpretation of DNA binding assays is further complicated by two issues: (1) there are no pre-defined

regions which one can quantify (akin to annotated genes in RNA-Seq). A common practice is to bin the genome in equal-sized bins or use peaks obtained from peak finders and (2) there is no direct conceptual correlate to the levels obtained from such assays which is similar to transcript abundance in RNA-Seq. Even with spike-ins, how can one interpret finding different amounts of the assayed protein in a given genome location between two different conditions? This can mean anything from the protein is bound in more/less cells, is more/less frequently bound or more/less stably bound...etc. Advances in both single-cell assays and spike-in technologies could eventually lead to accurate quantification and more transparent interpretations of such quantities.

2.5 Finding Patterns in Time-course Data

In order to analyze and integrate multiple data sets measuring multiple features over multiple time points, we need to develop suitable probabilistic models for integrating this data. Once we have obtained properly normalized data (see above), one can look for change patterns that occur in the data over time. Methods aiming to discover such time-course patterns can be grouped into three broad categories: (1) differential expression methods, (2) time-course clustering methods and (3) regulatory network inference methods.

2.5.1 Differential Expression Analysis

Differential expression analysis [88, 89] relies on statistical hypothesis testing. For each gene g, we want to carry out a hypothesis test with the null hypothesis being that the value of any gene g in condition $C1$ is not different from its value in condition $C2$, given the variation that we would expect by chance between replicate experiments. Therefore, this analysis requires having at least two replicate experiments for each condition and the output is typically a p value that expresses the probability of observing the difference between $C1$ and $C2$ in gene g by chance due to variation between replicate experiments. One main component to this approach when applied to high-throughput sequencing data is a statistical model of read counts. A popular model is the negative binomial distribution, which allows for a dispersion parameter that is independent of the mean. Because the number of biological replicates is usually too small to reliably estimate the variance of each gene independently, a key popular idea is to pool together genes with similar expression in order to obtain a more reliable estimate of the variance [90–92]. The exact models and statistical tests vary greatly and many differential expression programs have been developed, and benchmarked in [93] and [94]. Currently popular programs include edgeR [95] and DEseq [77].

In ChIP-Seq, DNase-Seq and ATAC-Seq, the same idea can be employed (often called “differential peak finding”). ChIP-Seq peaks and counts of reads in those peaks can be obtained from all replicates in both conditions and any of the differential expression programs can then be used to determine the peaks with statistically significant differences between the two conditions. Many automated differential peak finders have

been developed, and benchmarked in [96], usually based on the same ideas previously developed for gene expression data. Commonly cited programs include DiffReps [97] and PePr [98].

Of note regarding differential analysis is that it can err on the side of calling a gene not different when it really is (type II error). In the case of ChIP-Seq and DNase-Seq/ATAC-Seq, an alternative is intersecting peaks obtained independently from each condition; a genomic location which has a peak in only one condition is "differential". In this approach, one typically obtains more "differential" peaks than with statistically-rigorous "differential expression" analysis, which means more false positives from the point of view of the null hypothesis explained above, but less false negatives. Ideally, the two approaches should agree if the samples in both conditions were sequenced to saturation and normalized properly but this is almost never the case. One should attempt to apply as many of these approaches as possible to the data and check to see if and how conclusions might change.

2.5.2 Time-course Clustering Methods

Although there have been attempts to generalize the idea of differential expression to more than two conditions (see [99, 100] for examples), it is generally not possible to apply differential expression analysis when one has multiple covariates, like multiple histone modifications, across two or more conditions like multiple time-points. Hence, more integrative analysis has typically relied on clustering approaches, where one asks what (and how many) groups of genes exist in the data based on their time-course expression dynamics.

Clustering Time-course Gene Expression Data

Because of the availability of time-course microarray-based measurements of up to thousands of genes in various biological systems as early as the 1990s, modeling gene expression dynamics was one of the earliest problems in the analysis of high-throughput time-course data. The main general idea is to group genes based on their time-course expression profiles, such that each cluster contains a group of genes that share a similar time-course expression profile.

The earliest examples of gene clustering based on time-course gene expression profiles feature well-known clustering algorithms like simple Euclidean distance similarity matching [101], k-means [102], hierarchical clustering [103, 104], self-organizing maps [105], support vector machines [106], graph-theoretic approaches [107] and Gaussian mixture models [108]. The main drawback with such approaches is that the time-order of gene expression sampling is not taken into account. For example, when using k-means the outcome will not change if the user shuffles the order of the time-points. This problem was addressed first using three main methods: modeling of time-course gene expression using linear combinations of spline functions [109] which takes into account variable durations between sampled time-points, modeling time-course expression with auto-regressive functions [110] and the use of Hidden

Markov Models [111, 112], which directly encode time-course dependencies between the expression of gene g at time t_{i+1} and time t_i via the transition probability constraints of HMMs.

When grouping genes by their time-course behaviors, four main behaviors can be expected: "no-change", "non-monotonic inconsistent" changes, monotonic increase or decrease and cyclic profiles. Since cyclic gene expression profiles are especially expected in cell cycle studies, some algorithms were developed to discover such genes specifically [113, 114]. Other issues arise from non-uniform sampling of time-points and from low number of sampling points across time, which were addressed in [109, 115, 116] and [117, 118] respectively. Further, to differentiate the "no-change" group from "non-monotonic inconsistent" group, it was suggested to provide a "noise cluster" which is fixed and not learned from data, and to which genes showing inconsistent profiles or which do not belong clearly to one of the other clusters can be assigned [111].

Many such early algorithms and models used to cluster high- and mid-throughput time-course data are reviewed in [119, 120] and benchmarked in [121]. They remain instructive and often useful in terms of developing more complicated models and analysis of larger datasets. More recently, approaches designed strictly for clustering time-course gene expression have fallen out of fashion in favor of either differential expression analysis models (see above) or more integrative models for building gene regulatory networks (see below). This is partly due to the advent of complementary mid- and high-throughput transcription factor binding data and chromatin structure data which allow for meaningful inference of potentially causal relationships between genes.

Clustering Time-course Histone Modification Data

Clustering histone modification data across time complicates time-course clustering models because it invloves clustering multiple features (ie. histone modifications) across multiple time-points. Standard clustering methods like k-means and principal component analysis (PCA) can be used (see [122–124] for examples). However, as mentioned above such approaches ignore the time ordering of the data points. Generative models like Hidden Markov Models (HMMs) and more general Bayesian Network structures can be used to specify more sophisticated models that can take the time order of the data into account. A recent example, GATE [125], is a finite mixture of HMMs which can cluster multiple histone modifications over multiple time-points, taking advantage of the time order of the data. The genome is binned into equal-sized bins and time-course histone modifications in those bins are used to fit a multivariate Gaussian HMM with two states (active and inactive). All the HMMs from all the genome bins are clustered into a finite set of HMM classes (ie. time-course chromatin states). This model is limited mainly by the coercion of the chromatin state of a genome location to be in one of only two states [125], and that it is not possible to generalize it to hierarchical differentiation trajectories. Therefore, a more general model where there is no restriction on the complexity of the chromatin state trajectory or the cell stage lineage tree is desirable. In Chapter 6, I will describe a model that satisfies these criteria.

2.5.3 Regulatory Network Inference Methods

The goal of assigning enhancers to promoters using time-course data can be thought of as inference of gene regulatory networks (see Chapter 1). Historically such models, like time-course clustering (see above), started with the availability of high-throughput gene expression data using microarray technology. For example, Bayesian Networks were used to learn a causal network of genes based on time-course gene expression data [126]. A main distinguishing feature of the idea of building gene regulatory networks is that it is not assigning each gene to a single cluster. Unlike clustering of time-course gene expression data or clustering of time-course histone modification data (see above and Chapters 6 and 7), gene regulatory networks provide an interpretation of how genes relate to each other rather than mere clusters based on behavior. Constructing gene regulatory networks from time-course gene expression data has been an extremely popular area of research with what could be hundreds of models, ideas and variations published over the past two decades (reviewed in [127–129]).

With the availability of transcription factor binding data in the form of ChIP-ChIP (microarray-based) and ChIP-Seq, it became possible to build more elaborate models that model time-course gene expression data in terms of transcription factor binding [130–132]. One of the more prominent of such attempts is a model called DREM [133], which employs an input-output Hidden Markov Model to infer which transcription factor binding events regulate which change in gene expression dynamics of which genes. As such, each gene is assigned to potentially multiple transcription factor binding events that reflect its time-course gene expression dynamics. One main limitation of this model is that it takes into account dynamic gene expression data but not dynamic transcription factor binding data, assuming the binding landscape is static. This assumption holds if a certain transcription factor, given a certain chromatin context, will bind a specific set of sites and will not change its binding. This assumption, as we will show later in Chapter 7, is not necessarily always true. Further, DREM does not take into account the activation dynamics of enhancers. The same factor might not activate all regulatory regions it binds at the same rate. The rate might depend on transcription factor binding stability at a particular site [50], as well as the local chromatin environment. In Chapters 7 and 8, we discuss regulatory network inference in the context of enhancer-promoter interactions and the local chromatin environment.

Chapter 3

Introduction to Chromatin Regulation

3.1 Introduction

This chapter provides a summary of current ideas concerning chromatin regulation and how it relates to cell differentiation and development. It starts with a brief introduction to chromatin regulation followed by an introduction to differentiation and development model systems. We then discuss transcription initiation and chromatin structure and end with a discussion of promoter and enhancer chromatin dynamics during development and cell differentiation.

3.2 The *Epi*-genome: An Overview

In Eukaryotes, double-stranded DNA is packaged inside the nucleus with the help histone proteins forming "chromatin", which is folded in an intricate 3D structure. Nucleosomes, the most basic structural units of chromatin, are octamers of histone proteins around which 147 basepairs of DNA are wound [5, 6]. At this level, chromatin can be divided into "open chromatin" and "closed chromatin". Closed chromatin is nucleosome-rich and largely inaccessible to transcription factors and other transcription cofactors. Open chromatin is depleted from nucleosomes and is accessible to transcription factors and cofactors [134–136]. Open chromatin regions can be assayed using enzymes that cleave accessible DNA like DNase-I (see Chapter 2) and were historically, and correctly, associated with active transcription [137–139]. The idea of transcription regulation via closed and open chromatin gives the first example of *epigenomic* regulation where factors not necessarily related to DNA sequence such as histones and nucleosome positions can affect transcription and cell function (see [140] for a review).

On a higher level, chromatin is organized into topologically associated domains (TADs, see [141] for a review), whose boundaries are demarcated by CTCF binding sites. It is thought that enhancers can only access promoters that are within the same TAD, therefore TAD borders play a crucial role in development and differentiation (see [142] for an example). This indicates a different aspect of the epigenome where

factors like CTCF determine which enhancers can access which promoters and define the possible regulatory logic of each promoter (see [143] for a review).

Nucleosomal histones can also be covalently modified either on their core or on their tails which extend into the nuclear space [15]. Histone hyperacetylation has been long associated with actively transcribed, DNase-I hypersensitive, open chromatin and vice versa for histone hypoacetylation [144, 145]. Today, there is an ever growing list of possible histone modifications and possible combinations of histone modifications on the same nucleosome or the same genome region [9, 146]. Histone modifications can regulate chromatin structure either via crosstalk amongst each other in a way similar to signaling networks function or via changing the affinity of DNA to nucleosomes (see [147, 148] for examples).

Histone modifications are set and deleted via a large group of enzymes and complexes (reviewed in [149]). The importance of histone modifications was appreciated when it was realized that many known transcription activation cofactors are in fact histone modifiers (see [150] for an example). We now know that those complexes, as well as the histone modifications themselves, can also interact directly with transcription factors and the polymerase initiation and transcription machinery (see below).

Histone tail modifications have been the subject of extensive study and exploration in the past two decades aided by new genome-wide assays (see Chapter 2). This thesis is concerned with histone tail modifications and chromatin accessibility dynamics during differentiation. Other aspects of the epigenome like DNA methylation and histone core modifications are discussed in [151] and [152] respectively.

3.3 Development and Differentiation Model Systems

After the single-celled zygote is formed, the zygotic genome is first inactive and cellular processes are controlled by maternal RNAs and proteins. Zygotic genome activation is a key important event for embryogenesis that occurs over different time spans across different animals [153]. Another important event in embryogenesis is the gradual programming of pluripotency in the dividing zygote going through the totipotent Morula at the 16-cell change and the pluripotent Inner Cell Mass (ICM) inside the Blastocyst ([154], Figure 3.1). ICM gives rise to two layers of cells one of which, the Epiblast, goes through gastrulation and differentiates into the three germ layers: endoderm, ectoderm and mesoderm. Those three types of cells then gradually differentiate during embryogenesis into the multitude of different specialized cell types that occur in the adult organism. Model systems of embryogenesis and cell differentiation can be grouped into three broad categories: (1) *in vivo* embryogenesis in model organisms, (2) *in vitro* stepwise differentiation systems and (3) *in vitro* direct programming systems.

3.3.1 Embryogenesis and Stepwise Differentiation

Studying embryogenesis on the molecular level can be done in model organisms like worms, fruit flies, chiken, zebrafish and mice. However, studying chromatin dynamics

during *in vivo* animal development is technically challenging due to difficulties isolating a sufficient number of homogeneous cells following a defined lineage. Instead, many researchers rely on *in vitro* cell culture models of cell differentiation.

In *in vitro* models of cell differentiation, Embryonic Stem Cells (ESCs, reviewed in [156]) are often the starting cell fate (Figure 3.1). These cells were originally established from the ICM of mouse and human blastocysts in the 1980s and 1990s. ESCs have the ability to continuously proliferate in cell culture and can be induced to differentiate using various types of cues into specialized cell fates from all three germ layers, passing through stepwise stages similar to those observed during development (we call this here "stepwise differentiation", reviewed in [157]). Stepwise differentiation usually relies on manipulating signaling pathways by culturing the cells in media containing certain signaling molecules. In Figure 3.1, an example of stepwise differentiation is shown where gut tube cells are differentiated from human ESCs using different signaling molecules at different stages to recapitulate the *in vivo* differentiation stages of this cell lineage [158].

3.3.2 Direct Programming of Cell Fates

The study of histone modification dynamics in step-wise *in vitro* differentiation models is generally hampered by the fact that many such differentiation protocols are inefficient, meaning that a small percentage of cells successfully reach the intended terminal cell fate [159], potentially due to the induction of unintended gene regulatory networks [159]. Furthermore, cells are often not synchronized, meaning that there is no guarantee that all cells would make the same gene expression and local chromatin transitions at the same time. This heterogeneity makes it difficult to make inferences about promoter and enhancer chromatin dynamics from high-throughput cell population based assays like RNA-seq, ChIP-Seq, DNase-Seq and ATAC-Seq.

One system that overcomes such difficulties is the directed programming of mouse ESCs to post-mitotic spinal motor neurons (sMN) within 48 hours via the ectopic expression of three transcription factors: Ngn2, Isl1 and Lhx3 (NIL) [160] (Figure 3.1). In this system, ESCs harboring a Doxycycline-inducible copy of the NIL factors are first differentiated to embryoid bodies (EBs, three-dimensional pluripotent cell aggregates recapitulating many aspects of embryonic development [161]) by culturing the cells in suspension in synthetic defined media. NIL factors expression is then induced for 48 hours by the addition of Doxycycline to the media. NIL induction results in the successful differentiation of more than 90 % of the cells [160] by going through a fairly homogeneous differentiation process, as confirmed by single-cell RNA-Seq [155].

This NIL system is distinguished from stepwise differentiation protocols by the fact that pluripotent cells are directly converted to motor neurons by the ectopic expression of a specific set of transcription factors without passing through the developmental stages of motor neuron differentiation [160]. This system is an example of direct conversion between cell fates using ectopic induction of transcription factors expression, often called "trans-differentiation" or "reprogramming" or "direct programming". This

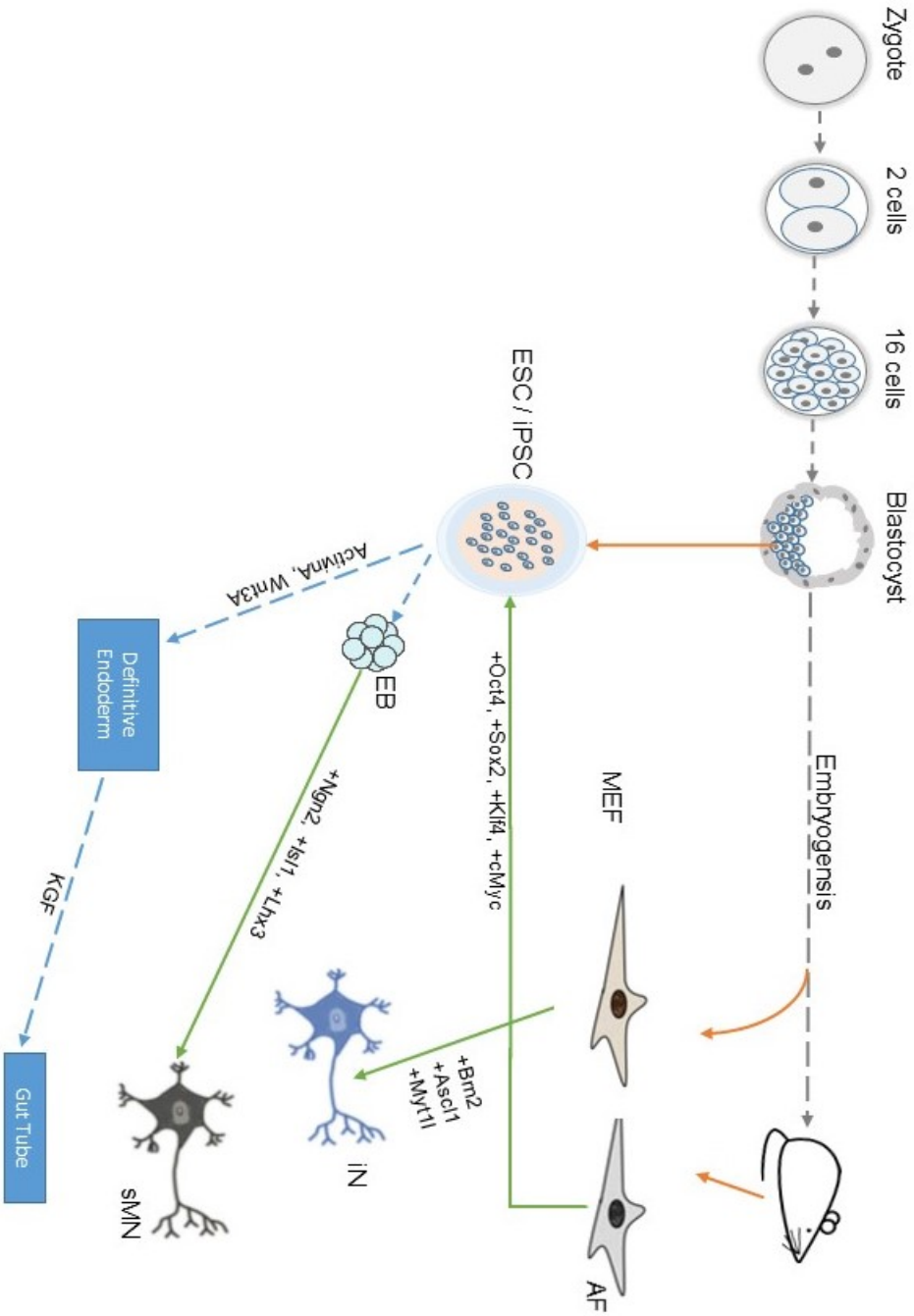


Figure 3.1: Embryogensis starts with the single-celled zygote going through the blastocyst containing the Inner Cell Mass (gray dashed edges). Embryonic Stem Cells (ESCs) are isolated from the Blastocyst Inner Cell Mass. They can be differentiated *in vitro* to cell types from all three germ layers in a stepwise fashion (blue dashed edges). Mouse embryonic (MEFs) and adult Fibroblasts (AFs) can be directly reprogrammed back to induced pluripotent stem cells (iPSCs) via ectopic expression of pluripotency transcription factors. MEFs can be directly programmed to induced Neurons (iN) via expression of the BAM factors. ESCs can be differentiated to Embryonic bodies (EBs) which can be directly programmed to spinal motor neurons (sMN) by ectopic expression of the NIL factors. Direct programming is indicated using solid green arrows. Neuron and EB picutres were adapted from [155]. Mouse cartoon is by Seans Potato Business (Own work) [CC BY-SA 3.0], via Wikimedia Commons.

strategy has been used to interconvert directly between many different cell fates (reviewed in [162] and [163]), albeit with low differentiation efficiency [159]. The most prominent example of direct programming is the conversion of embryonic and adult mouse fibroblasts to induced pluripotent stem cells (iPSCs) via ectopic expression of the "Yamanaka Factors" (Oct4, Klf4, Sox2 and cMyc) [164] (Figure 3.1). iPSCs are pluripotent and have many ESC properties and can contribute to embryogenesis if injected in the Blastocyst [164]. Conversion of fibroblasts to iPSCs was a major milestone in direct programming of cell fates because it "reprogrammed" a cell of reduced potency (fibroblast) to one that is pluripotent (iPSC).

Another system of direct programming is the conversion of mouse fibroblasts to so called induced neurons (iN) by ectopic expression of the BAM factors (Brn2, Ascl1 and Myt1l, see Figure 3.1, [165]). This system is also interesting because it directly "trans-differentiates" between two different cell types that are not known to be accessible from each other during normal embryogenesis and development.

The low efficiency of most direct programming protocols, including the Yamanaka Factors system and the BAM system, obfuscate the interpretation of the results. The reasons for differentiation low efficiency is a current area of research [159, 166, 167] but it can be attributed to the fact that the induced transcription factors are chosen based on experience with the desired cell type regulatory network. One goal of the stem cell community is to be able to *rationally design* transcription factor mixes that can efficiently convert one cell type into another. In recent concurrent work, [167] dissect the inefficient BAM fibroblast-to-iN conversion system using single-cell RNA-Seq data and find that this trans-differentiation system is inefficient due to the induction of unintended myogenic gene regulatory network, which is consistent with previous analysis of other systems [159].

However, when one wants to understand how a certain system works in order to engineer similar systems, it is desirable to reverse-engineer a system that works efficiently rather than one that works inefficiently. In Chapters 6 and 7, we take advantage of the efficient NIL differentiation system in collaboration with the Mazzoni lab at New York University to investigate promoter and enhancer chromatin dynamics during motor neuron direct programming [155].

3.4 Chromatin Regulation and Transcription Initiation

Gene transcription starts by the assembly of the Pre-initiation complex (PIC) at core promoters, defined as 100bp stretch of DNA surrounding TSSs. PIC is composed of Polymerase II (PolII, the enzyme that catalyzes transcription), TBP (TATA Binding Protein) and TAFs (TBP Associated Factors). Although the PIC is not as sequence specific as transcription factors, it does display certain affinities to certain sequence elements such as the TATA box, Initiator motif and the DRE motif [168–170]. Different core promoters can include different combinations of these core promoter motifs [171]. In fact, the transcription initiation machinery and sequence motifs were not only greatly diversified across evolution [172, 173], but even in the same organism different

combinations of core promoter motif combinations are potentially engaged by different variants of the PIC (which include different combinations of TAFs), contributing to different specificities of promoter-enhancer interactions [169, 174–181]. This poises the transcription initiation process by itself as an important regulator of gene activation and repression.

Transcription initiation in eukaryotes occurs in a discontinuous manner showing discrete bursts of transcription initiation with intermittent periods of transcription silence [182–186], which introduces two main parameters: the burst frequency and the burst size (how many transcripts per burst). Variations in these two parameters is attributed to the diversity of the core promoter motifs across different promoter regions and the promoter chromatin environment [187–189]. Promoter chromatin in most, if not all, eukaryotes is marked by "nucleosome-free regions (NFRs) which allow physical access to the DNA thereby restricting [the PIC complex] to relevant places [[190, 191]]" [8]. "These regions are typically flanked by well-positioned nucleosomes" [8] downstream of the TSS, and often upstream of the TSS as well [191–196]. The well-positioned nucleosomes flanking promoter regions can be considered the "canonical view" of promoter chromatin. An alternative view concerns promoter regions with less well positioned nucleosomes; meaning that the promoter region is not kept in a bona fide nucleosome-free state but that nucleosomes are potentially "statistically positioned" over the promoter region in a stochastic manner [197, 198]. This leads to a view of two different strategies for promoter chromatin regulation that appear to be common to many eukaryotes: (1) promoter regions with less well-positioned nucleosomes which tend to have more variable gene expression response, clearer/stronger core promoter motifs and focused transcription initiation from one or very few closely-spaced TSSs and (2) promoter regions with well-positioned nucleosomes and canonically open region which tend to have "weaker" core promoter motifs and more dispersed transcription initiation patterns [197, 199–203]. Such strategies, intimately linked to promoter chromatin architecture, have important implications on promoters' response to stimuli and how their sequence and gene expression patterns evolve [196, 197, 204–207].

Another aspect of promoter chromatin regulation is the covalent modifications of the histones in the nucleosomes flanking the promoter NFR. "Nucleosomes in the downstream gene body adjacent to [TSSs] are marked by a combination of histone [modifications] including acetylation at lysine 27 and lysine 9 of histone 3 (H3K27ac and H3K9ac), [as well as] a cascade of spatially organized H3K4 methylation states with trimethylation being the most promoter-proximal, moving to dimethylation and then to mono-methylation as one moves down the gene [124, 193, 208–212]"[8]. "Interplay between [the COMPASS family of protein complexes (which are responsible for H3K4 methylation),] the Paf1 transcription elongation complex and phosphorylation of the C-terminal domain (CTD) of PolII, is thought to dictate di- and trimethylation of H3K4 around active promoters [213]. [On the other hand,] trimethylation of H3K4 has been shown to influence initiation at promoters via interactions with TAF3 [214, 215]" [8]. In addition, members of the transcription initiation machinery were shown to have histone acetyltransferase activity [216], while the acetylation of H3K27

itself was shown to correlate with the release of PolII into elongation [217]. These examples demonstrate the important links between transcription initiation and histone modifications. Indeed, histone modification levels were shown to be predictive of gene expression (whole cell polyA-selected mRNA levels) [218], of polymerase ChIP-Seq levels [219] and of nascently transcribed RNA levels [Maja Schuster / Ohler Lab, Not Published].

3.5 Promoter Chromatin State Dynamics during Development

Studying promoters' roles during development and differentiation requires a dynamic view of promoter regulation. A common theme in promoters' role in gene expression control is the use of alternative promoters and TSSs [220]. Transcription initiation dynamics at single basepair resolution were mapped at 1 hour intervals during the life cycle of *D. melanogaster* [221], where more than 40 % of genes were found to have at least two alternative promoters. Alternative promoters from the same gene were on average not correlated over time meaning that core promoters tend to be differentially used at different times during fly's lifecycle. In zebrafish, two studies showed that during the maternal to zygotic transition nucleosome arrays at promoters appear [222] and a switch occurs from TSSs with AT-rich core promoter sequences with poorly positioned nucleosomes to TSSs with weaker core promoter sequences but more precisely positioned nucleosomes [223]. The same two themes of promoter architecture were reproducibly observed in many organisms and linked to developmentally regulated genes (see above), but it is not clear whether alternative promoter usage during development is also common to mammals.

Different promoters can exhibit different rates of response to activation which was shown to be modulated by the frequency of transcriptional bursts while keeping burst sizes constant [224], which in turn might be related to promoter chromatin dynamics [223]. Indeed, the competition between the transcription initiation machinery and nucleosomes influences promoter dynamics during development [225]. Therefore, it is likely the use of alternative promoters is related to differential accessibility of promoters to the transcription initiation machinery. But what role do histone modifications play in this?

During programming of pluripotency from the single-celled zygote, histone modifications undergo global dynamic changes [227] as well as local specific changes that are required for the establishment of pluripotency (see [228] for an example and [154] for a review). Once pluripotency is established, a unique chromatin state can be observed at promoters in ESCs, and other *in vitro* and *in vivo* [229] pluripotent cells, called "bivalent chromatin state" (reviewed in [226]). Bivalent promoters are defined by the co-presence of both activating H3K4me3 modification and repressive H3K27me3 modification [230, 231] (Figure 3.2), often asymmetrically on the same nucleosomes [232]. with promoters of developmental and lineage-specific genes are typically in this bivalent state [30, 230, 233, 234]. Establishment of bivalent chromatin is linked

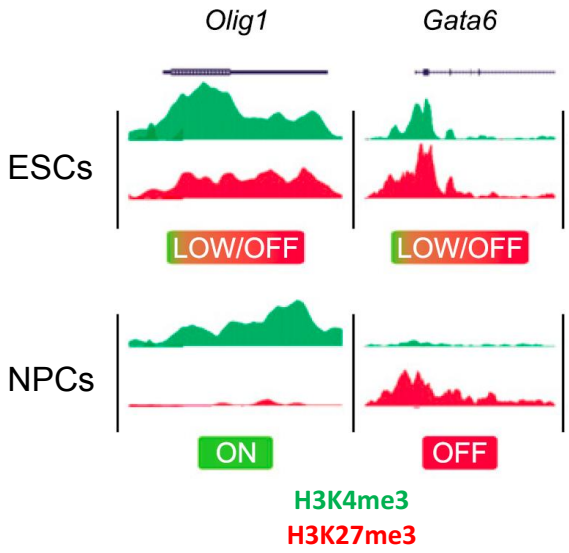


Figure 3.2: Examples of genes that are bivalent at ESCs and are resolved when cells differentiation to Neural Precursor Cells (NPCs). Figure adapted from [226].

GC-rich promoters which are associated with H3K27me3 in stems cells [235] and are thought to be able to recruit the Polycomb Repression Complex 2 (PRC2) complex [236]. EZH2, a subunit of the PRC2 also co-localized with developmental genes in ESCs [237], is responsible for setting the tri-methyl modification on H3K27. A likely mechanism for repression of bivalent promoters is that H3K27me3 provides a substrate for Polycomb PRC1 complex binding [238, 239] which then inhibits polymerase initiation [240] and elongation [241, 242]. Indeed, bivalent promoters bound by PRC2 only show polymerase pausing but no or decreased elongation [243], while promoters bound by both PRC1 and PRC2 show decreased pausing [243] and are bound by unproductive form of Polymerase II [244]. It should still be noted though that at least a subset of bivalent promoters do produce functional RNA species including mRNA [234, 241, 244] and short non-coding RNA [245]. However, this seems to be independent of PRC complexes [245] but rather related to stochastic switches to an active promoter state [244]. Interestingly, while polycomb silencing was not found to be important for stem cell maintenance, it is important for proper differentiation [246]. All these observations combined lend support to the prominent idea that bivalent promoters are "poised" for activation [237, 247] and potentially facilitate differentiation by exhibiting a lower threshold and more timely response to activation signals compared to promoters repressed via other mechanisms [226].

Activation of bivalent promoters during differentiation is marked by loss of H3K27 methylation and further gain of H3K4me3 [30, 233] (Figure 3.2). The Trithorax group of proteins, specifically Mll2 [248], are responsible for depositing H3K4 methylation in bivalent promoters. H3K4me3 interacts with the transcription initiation machinery

[214, 215] and its presence can antagonize deposition of H3K27me3 [232, 249]. Active promoters in terminally differentiated cells are also known to feature H3K27ac in promoter-proximal nucleosomes (see above). The depletion of H3K27me3 from bivalent promoters during their activation implies a switch from methylation to acetylation on H3K27 [250, 251]. The Trithorax protein group was also shown to interact with CBP, an H3K27 acetylase, to promote acetylation and to antagonize H3K27 methylation [252], while PRC1 itself was shown to inhibit acetylation of H3K27 [253]. Further, knockdown of Suz12 (part of PRC2) results in increased levels of H3K27ac at promoters and increased gene expression levels from these promoters [250].

These studies paint a complicated picture of bivalency resolution that involves interactions and feedback loops between the PRC complexes, the Trithorax group of proteins, histone acetyltransferases and the transcription initiation machinery. The dynamics of these interactions during differentiation and their relationship to the responsiveness of promoters to activation signals are not yet understood. In one study, bivalency resolution dynamics were investigated during step-wise differentiation of ESCs to cardiomyocytes [122] via time-course ChIP-Seq. Developmental genes were in a bivalent state in the pluripotency stage. During differentiation, promoters relevant to to cardiomyocyte fate showed a slow gradual increase in H3K4me3 and parallel gradual decrease in H3K27me3. Genes not relevant to the cardiomyocyte lineage mostly showed a decrease in H3K4me3 suggesting a resolution of bivalency towards a *bona fide* repressed state [122]. However, this study did not investigate the role of H3K27 acetylation in resolving bivalency and suffered from inconsistencies in the differentiation efficiency at the different time-points limiting the power of observations based on cell population-based ChIP-Seq [122].

3.6 Enhancers Chromatin Dynamics during Development

Transcription enhancers are the most characterized class of distal regulatory elements [10, 11, 13]. Enhancers contain DNA sequence motifs which act as binding sites for transcription factors (see Chapter 1). Transcription factor binding to enhancers regulates promoters activation potentially through physical proximity to their target promoters in three-dimensional space. Enhancers are thought to be able to act on their target promoters regardless of orientation, exhibit degeneracy in their promoter specificities [11–13, 254, 255] and might operate by regulating transcription bursts from promoters [256–258] (see above).

Pluripotent cells are characterized by a permissive, highly mobile chromatin architecture. They feature a higher number of accessible distal regulatory elements than in differentiated cells [259–261], a highly mobile chromatin architecture [262, 263] and they are nearly devoid of constitutive heterochromatin histone modifications like H3K9me3 [264]. At the early 2-cell stage, the zygotic genome is characterized by large permissive open regions that are then gradually refined into distinct well-defined open regions demarcating the locations of distal and proximal regulatory elements [261].

The number of open proximal and distal regulatory regions then steadily increases up to the Morula totipotent stage [259]. These observations can explain the high number of open regions observed in pluripotent cells like ESCs [260]. During differentiation, the enhancer landscape starts to get limited gradually by the appearance of large repressed heterochromatin domains [264] and the loss of a large number of open regions [260].

Therefore, the available enhancer landscape is gradually restricted during differentiation to reflect a particular cell fate. In fact, it is widely established that during development and differentiation tissue-specific enhancers orchestrate cell function and the required gene expression program (see [265–268] for examples and [269] for a review). One key idea behind enhancer tissue-specificity is that each enhancer will contain motifs for multiple factors potentially arranged under certain spatial constraints, which would allow for transcription factor cooperativity at a single enhancer (see [270, 271] for examples and reviews). The combination of transcription factors binding an enhancer, how the sequence motifs of those factors are arranged in this enhancer and how well they match the factors sequence preferences [272–274] would determine whether the enhancer becomes active (see [271] for a review). The view of enhancer regulation becomes more complicated when one considers that a single promoter can be regulated by the action of multiple enhancers often cooperating to confer complex enhancer-promoter activation dynamics that are not yet fully understood (see [275–279] for examples).

Enhancer dynamics, like promoter dynamics, can be characterized using time-course histone modification data in *in vivo* and *in vitro* differentiation systems. Enhancer activation can occur with the help of "Pioneer" transcription factors (discussed in [280]), which are thought to activate nucleosome-rich sites (ie. closed chromatin) by binding to parial motifs exposed on nucleosome surfaces [281]. In several pioneering studies, inactive / poised enhancers were found to be closed (ie. not nucleosome-free) and marked with monomethylation on lysine-4 of histone 3 (H3K4me1). H3K4me1 was also shown to be a reliable mark for discriminating subsequently activated enhancers from those that will not be activated later during differentiation [158]. When they are activated, enhancers become nucleosome-free and accessible to further transcription factor binding with their flanking nucleosomes gaining further methlyation in the form of H3K4me2 and acetylation in the form of H3K27ac ([282–286], Figure 3.3). Enhancer transcription is thought to occur after transcription factor binding and is associated with gaining H3K4me2 [287].

Identification of enhancers in a genome-wide manner has typically relied on high-throughput sequencing data such as DNase-Seq / ATAC-Seq where nucleosome-free regions that do not coincide with promoters are thought to be active enhancers [208], or transcription factors ChIP-Seq with which one can identify distal transcription factor binding sites [272], or using ChIP-Seq for histone acetylase P300 [288]. It was shown that enhancers in mammals also feature transcription initiation of short unstable non-coding RNAs called eRNA [289, 290] (see Chapter 5), which is another feature that has also been used to identify the enhancer landscape in various cell types [291, 292], although it is still controversial whether eRNA production is required or sufficient for

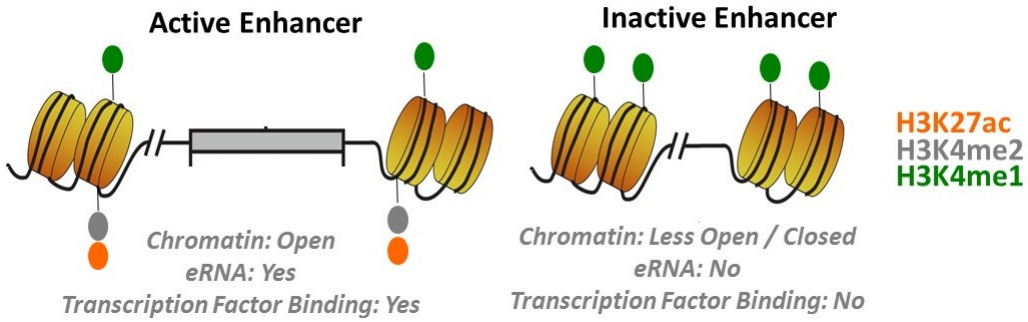


Figure 3.3: Chromatin Environment of Active and Inactive Enhancers. Figure adapted from [298]

enhancer activity [293]. Finally, high-throughput enhancer-reporter assays, in which genomic sequences are assayed for their ability to enhance transcription from a gene promoter, have been used to identify potential enhancer locations in a genome-wide manner [294–297].

To identify target promoters of an enhancer or a group of enhancers, researchers have turned to computational machine learning models that aim to link enhancers to promoters based on correlation of activity as measured by histone modifications and other features [191, 299, 300]. This means that enhancers can be linked to promoters using probabilitic models that attempt to predict gene expression from enhancer activity, often taking the distance between enhancers and promoters into account (see [301–303] for examples). These models rely on the fundamental concept that enhancers can be grouped into distinct classes based on their activation dynamics, as measured by histone modifications for example, and that each such class will regulate a certain distinct class of target promoters leading to a specific gene expression pattern (see Chapter 7).

Chapter 4

Joint Analysis of Genome-wide Sequencing Data

4.1 Contribution Statement

Results in this chapter are based on work in the following publications: [73] and [298].

Contributions to Chapter Results: Mahmoud M Ibrahim developed the peak finder (JAMM), performed all related benchmarking analysis and developed the chromatin state pipeline and performed its associated analysis. Scott A Lacadie (Ohler Lab / Max-Delbrueck-Center in Berlin) advised on the peak finder benchmarking.

4.2 Introduction

4.2.1 Peak Finding[1]

Peak finding involves separating the genome into regions of high enrichment (i.e. peaks or clusters or binding sites) and regions of low enrichment (see Chapter 2). However, most peak and cluster finding programs are developed with a specific experimental protocol or dataset type in mind [66, 72]. Therefore, it is usually difficult to apply the same analysis pipeline uniformly across all datasets in a given experiment.

Recently, there were attempts to develop *universal* peak finders by defining the problem as that of classical signal detection [72]. The main advantage of this approach is that it allows for uniform data analysis via theoretically proven optimal signal detectors. The main drawback is that it ignores the fact that for many protocols, enrichment sites in the same dataset are not expected to have the same signal properties. For example, DNase hypersensitive regions are expected to have different widths and signal-to-noise ratios (SNR) [302]. Therefore, there is a need for an approach that would not only focus on optimal detection of enrichment site locations but would also be able to adapt to enrichment sites with different signal properties in the same dataset.

[1] Text in this section is largely copied directly from [73]

Another drawback in current peak finding approaches is how biological replicates are processed. While others have focused on integration of multiple datasets to detect co-occurrence or differential enrichment (see [304], [97], [305], [72] and [306] for examples), common solutions for integrated replicates analysis are (1) detecting enrichment sites separately followed by combining the result towards "consensus" sites via union or intersection [307, 308] or (2) pooling the aligned reads from all replicates followed by detecting enriched sites (see [309] and [310] for examples). However, taking the intersect or union of separately detected sites mandates re-scoring the peaks and leads to inaccurate enriched sites widths. On the other hand, pooling alignments before site detection assumes that the underlying biology is perfectly reproducible so that pooling biological replicates is analogous to sequencing one experiment to higher depth. However, even if biological replicates are highly reproducible, pooling them may still lead to invalid spatial information and invalid sorting of peaks in genomic locations where they do not perfectly agree. Therefore, there is a need to develop a method for integrated analysis of biological replicates that takes advantage of the differential spatial and intensity information in the separate replicates.

In this chapter, I will introduce JAMM (Joint Analysis of NGS replicates via Mixture Model clustering); a universal peak finding pipeline that can integrate information from multiple biological replicates and adapt to different enriched sites signal properties even if in the same dataset. We focus on ChIP-Seq [311], since most peak finders were developed and tested using this protocol. We compare several programs that focus on different aspects of the peak finding problem. MACS [65] models read counts using a local Poisson distribution to improve specificity (see Chapter 2), PeakRanger [312] focuses on detecting neighboring narrow peaks at high resolution, PeakZilla [70] is designed for uniform punctate ChIP sites like transcription factor binding sites from ChIP-exo [**exo**] and ChIP-Seq, BCP [68] develops explicit formulas to model read counts, CCAT [313] detects enrichment patterns with low SNR and DFilter [72] is a universal peak finder based on optimal signal detection.

4.2.2 Chromatin States

Peak finders can summarize only one ChIP-Seq or DNase-Seq data set at a time. To integrate information from multiple data sets, researchers have turned to unsupervised clustering approaches including k-means and HMMs in order to discover co-localized histone modification combinations, called "chromatin states", and to assign different regions of the genome to these states [212, 314–317]. The assignment process, called "genome segmentation", partitions the genome into non-overlapping segments which have different labels expressing the histone modification combinations in those genomic segments.

One of the most prominent approaches is ChromHMM [315, 316] which is a completely unsupervised HMM approach. ChromHMM starts with binarizing the signal of each histone modification (ie. peak calling) at 200 bp resolution using a simple procedure based on modeling read counts with a Poisson distribution. After binarizing the signal of each histone modification, an HMM is fit using a modified version

of the Baum-Welch algorithm [315]. This HMM features multivariate Bernoulli emissions: briefly, each histone modification in each state is modeled using a Bernoulli distribution that defines the probability of observing this histone modification at this state. Histone modifications in the same state are independent of each other. The final emission distributions learned can be interpreted as the frequency of observing each histone modification in a given state. Due to binarization ChromHMM does not take into account differences in read counts between multiple regions of the genome. Furthermore, ChromHMM considers different histone modifications to be independent of each other, an assumption which is clearly an oversimplification [318].

Another popular approach is Segway [317] which is not strictly an HMM but a more general Dynamic Bayesian Network (DBN) structure which attempts various improvements over ChromHMM. Segway starts by transforming ChIP-Seq read counts using the inverse hyperbolic sine function which can transform read counts to a Gaussian-like distribution. The Segway DBN model is then learned on the transformed read count data using the EM algorithm. The model assumes the transformed data follow a Gaussian distribution, therefore taking into account the differences in the ChIP-Seq signal between different genomic regions. However, the variance parameters are tied for each histone modification across all states and histone modifications are, like in ChromHMM, assumed to be independent of each other. Further, Segway operates on single basepair bins, giving it a potentially higher resolution than that of ChromHMM but making it significantly more computationally expensive. To ensure that Segway will not assign single basepair segments, a "duration" model is integrated into the Segway model that can be set to allow for a minimum segment length [317].

Concurrent with the work in this thesis, various other unsupervised chromatin state solutions were proposed. Those include a "bidirectional" HMM that has two sets of transition probabilities to express the two different strands of the genome [319], which makes it especially useful for integrating stranded RNA-type data with histone modification data [55]. Additionally, a very similar model to ChromHMM was proposed with the main difference being that the model parameters are learned using spectral methods instead of the Baum-Welch algorithm, which can potentially eliminate the class imbalance problem that learning models using EM-related algorithms typically suffer from [320]. Finally, an HMM model with discrete multivariate emission probability distribution was introduced for directly modeling the read counts. This model can also take dependencies between histone modifications into account and does not require preprocessing or transformation of the data [321].

In this chapter, I will introduce a chromatin state segmentation and discovery pipeline that can obtain chromatin states at high resolution (10 basepairs) taking into account dependencies between histone modifications and differences between histone modification signals in different regions, but without requiring expensive computational resources.

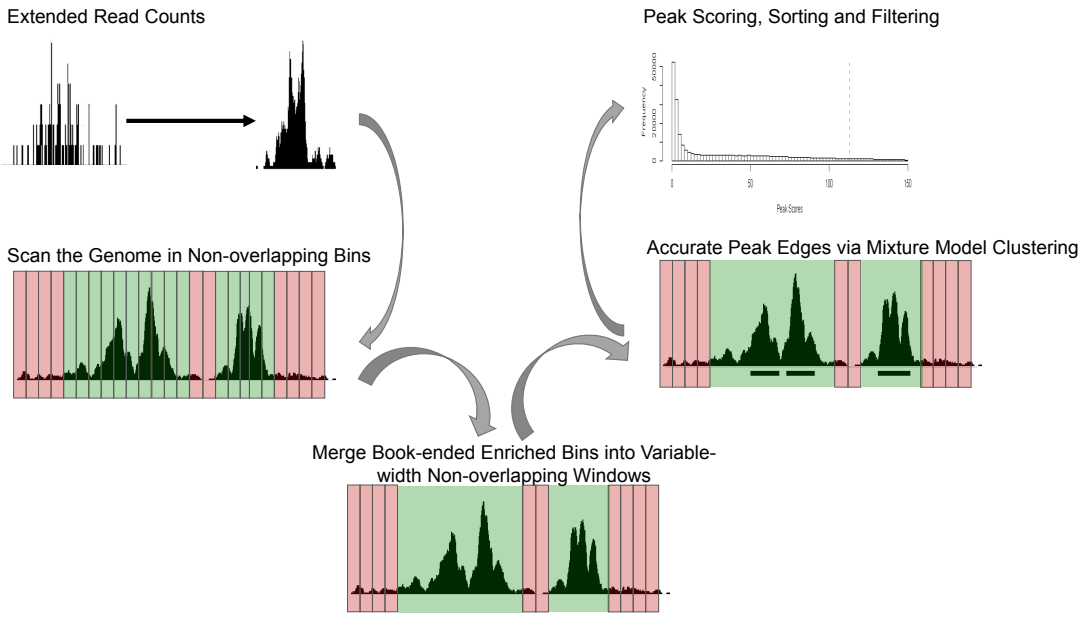


Figure 4.1: JAMM Peak Finding Steps

4.3 Results

4.3.1 Analysis of NGS Replicates via Mixture Model Clustering[2]

JAMM's core peak finding steps (Figure 4.1) involve selecting local variable-width windows that are enriched over background, followed by clustering the normalized extended-read counts in those windows into a peak cluster and noise cluster(s). Peak finding via local clustering allows JAMM to adapt to peaks with different shapes and signal properties and to accurately determine peak boundaries. Furthermore, using clustering as an approach for peak finding extends naturally to multivariate clustering, which is useful for integrating datasets that are similar but not exactly the same, such as replicates. We chose clustering via multivariate Gaussian mixture models, which is convenient for including information about the covariance of the biological replicates.

Selecting Enriched Windows

To be able to adapt to different peak shapes in the same dataset, JAMM selects enriched windows first and then assigns peaks locally in those windows. To find enriched windows, the genome is divided into small non-overlapping bins of equal widths and a decision is made whether each bin is enriched over background or not. All book-ended enriched bins are then merged into larger, non-overlapping, variable-width windows. This approach ensures that enriched windows include entire binding sites and

[2]Text and figures in this section are largely copied directly from [73]

that JAMM can seamlessly adapt to enrichment domains of different widths even if in the same data set.

Similar to [67], JAMM selects the bin size Δ that minimizes the cost function $C_n(\Delta)$ [322]:

$$C_n(\Delta) = \frac{2k - v}{(n\Delta)^2}$$

where n is the total number of reads, k is the average number of reads per bin for bins with width Δ and v is the variance. The estimated bin size is typically between 50 and 200bp. The user can also specify an arbitrary bin size.

A bin is enriched over background if $\mu_s > \mu_b$ where μ_s and μ_b are the average normalized extended-read counts in the sample bin and the corresponding background bin respectively.

Smoothing Read Counts

In order to produce an appropriate data vector for Gaussian mixture model clustering, JAMM smoothes the normalized extended-read counts in each enriched window using a two-pass ARMA (Auto-regressive Moving-average) filter. In the first pass, every basepair $bp(t)$ is determined by:

$$bp(t) = bp(t-1) + \sum_{i=0}^{a} bp(t-i) \times b$$

where a and b are 80 and 0.0125 respectively, by default. The second pass follows the same equation but in the reverse direction in order to correct for the phase-shift introduced by the first pass. In implementation, JAMM utilizes the function *filtfilt* in the R package *Signal* [323].

Assigning Accurate Peak Boundaries

JAMM assumes that the signal (smoothed extended-read counts) in enriched windows originated from a (multivariate) Gaussian mixture model [19]:

$$\prod_{t=1}^{T} \sum_{k=1}^{K} w_k \times N_k(bp_t | \mu_k, \Sigma_k)$$

where T is the window size, K is the number of components, bp_t is the read signal value for basepair t, w_k is the weight of component k in the mixture and μ_k and Σ_k are the vector of means and the covariance matrix for component k respectively.

To find peak locations in enriched windows, JAMM learns the Gaussian mixture model parameters using the Expectation-Maximization (EM) algorithm, assuming either 2 (corresponding to peaks and noise) or 3 mixture components (corresponding to peaks, peak tails and noise) [324, 325]. The mixture component with the highest mean is taken to be the enriched cluster, and contiguous basepairs assigned to this cluster

are taken to be the peaks. The covariance matrix is assumed to be different amongst different components and is parameterized according to its eigenvalue decomposition [19]:

$$\Sigma_k = \Upsilon_k \times \Lambda_k \times \Upsilon_k^{-1}$$

and

$$\Lambda_k = \lambda_k \times A_k$$

where Υ is the orthogonal matrix of the eigenvectors and Λ is a diagonal matrix with the eigenvalues at the diagonals, with λ_k being the first eigenvalue in Λ_k and A_k being a diagonal matrix with a vector at the diagonal that is proportional to the vector of eigenvalues. Therefore, Υ_k determines the orientation of the eigenvectors of k while λ_k defines the volume k occupies in the n-dimensional space and A_k defines the shape of the contour lines.

Peak Scoring

The background signal in every peak is subtracted from the corresponding sample signal. The resulting background-normalized signal values are averaged to produce the mean peak background-normalized signal (μ_{ns}). In addition, JAMM executes the Mann-Whitney-U non-parametric test to compare the sample signal (not background normalized) to the corresponding background signal. A Benjamini-Hochberg correction is applied to the full list of p-values, after peak finding is complete [326]. JAMM defines the peak score to be:

$$S_p = \mu_{ns} \times -log10(p_{corrected})$$

.

In this case, the p-value serves two purposes: it biases the overall peak score distribution in favor of peaks with bigger widths and for peaks with comparable widths it biases the peak score in favor of peaks whose signal is more statistically significantly different relative to background.

Peak Finding Accuracy

First, we sought to establish that JAMM achieves a similar or better site detection specificity compared to other recently published peak finders including MACS [65], BCP [68], PeakZilla [70], PeakRanger [312] and DFilter [72]. Specificity refers to the extent to which peak finders can determine the correct locations of enriched sites. Because there is no gold standard for benchmarking peak finders, we employed three different benchmarks focused on transcription factor ChIP-Seq: (1) motif finding precision (defined as fraction of peaks with motif matches out of all peaks called) using FIMO which utilizes a uniform zero-order background model [327], (2) motif likelihood (defined as the maximum motif likelihood obtained) using SpeakerScan which

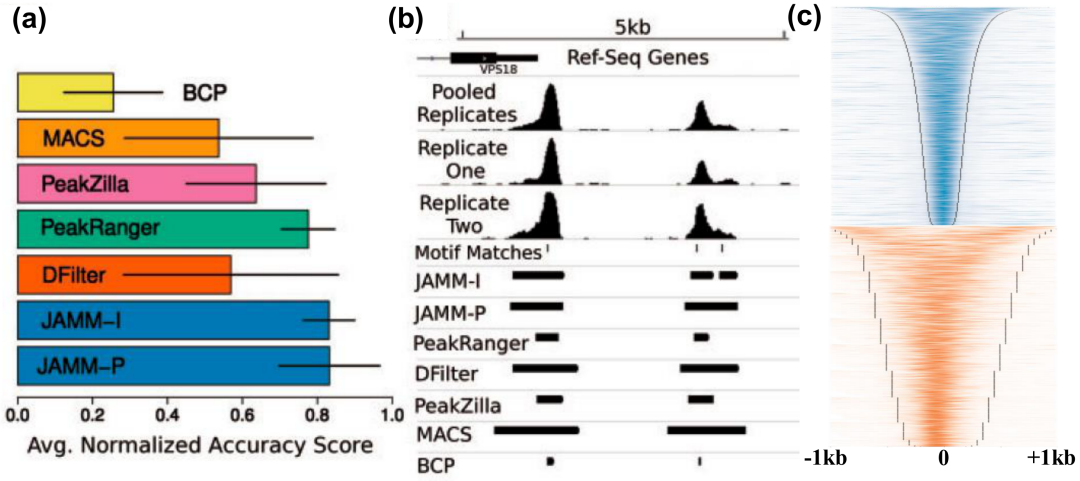


Figure 4.2: (a) Average normalized accuracy score over three benchmarks, see Appendix A. (b) An example of JAMM-Is improved spatial resolution because of replicate integration (CTCF, K562 from Encode [330]). (c) Peak width determination: heatmaps are centered on peak center, ranked by peak width and show extended-read count intensity and the corresponding peak edges (gray squares) for Encode HeLa-S3 DNase-Seq data (University of Washington) [330], using JAMM (top, blue) and DFilter (bottom, red). Figure adapted from [73].

utilizes a first-order local background model [328] and (3) recovery of manually curated positive peaks as reported in [329] (defined as the total number of peaks that intersected any manually curated positive peaks after subtracting the total number of peaks that intersected only manually curated negative peaks). When we consider the results over all datasets and all benchmarks, we find that JAMM and PeakRanger [312] are the top ranking programs (Figure 4.2a). JAMM ranked first for one benchmark (motif likelihood) and second for the other two benchmarks (see Appendix 1).

When comparing JAMM running on the replicates separately (JAMM-I) to JAMM running on pooled replicates (JAMM-P), we found that JAMM-I consistently outperforms JAMM-P (JAMM-P ranked better than JAMM-I in only one comparison) (Appendix 1), indicating that JAMM's replicate integration improves peak finding specificity over replicate pooling. A main contributing factor is JAMM-I's better spatial resolution due to replicate integration via multivariate mixture model clustering. Figure 4.2b provides a demonstration of JAMM-I's improvement over replicate pooling. Only JAMM-I can resolve two neighboring CTCF binding sites: the pooled replicate profile obscures the better spatial resolution of Replicate One due to the poorer resolution of Replicate Two.

Spatial Resolution

Open regions assayed by DNase-Seq and ATAC-Seq are expected to occur at variable widths throughout the genome [302]. Correlating DNase-Seq peak widths determined

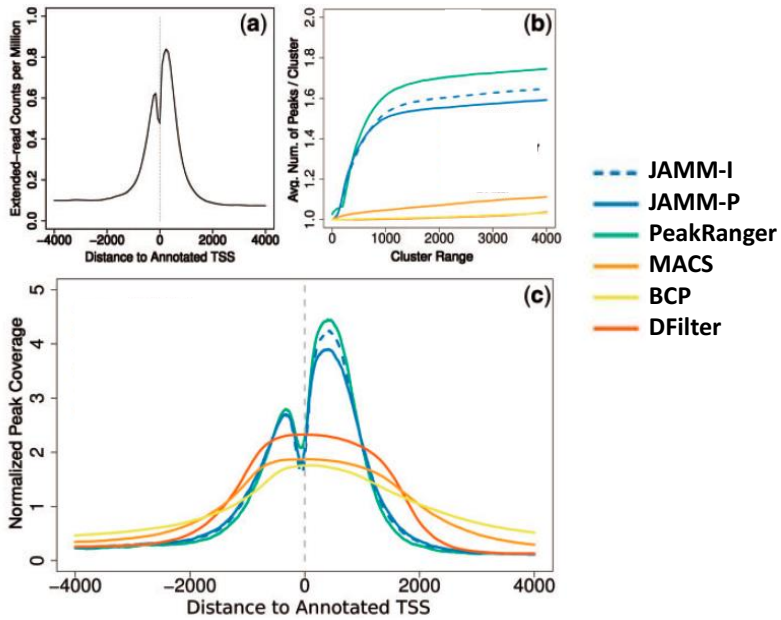


Figure 4.3: Only JAMM and PeakRanger can recover the resolution of the dataset (a) in the peaks called (b) and (c). (b) Shows the average number of peaks per cluster at different cluster ranges. Cluster range is the maximum distance separating peaks in the same cluster (for example, if two peaks are 50 bp apart, they will be grouped together in one cluster if cluster range is 50 bp or more). Figure adapted from [73].

by JAMM versus those determined by DFilter to the DNase-Seq read counts, one can observe that JAMM can assign peak boundaries corresponding accurately to variable-width open region boundaries while DFilter [72] can not (Figure 4.2c).

Spatial resolution is also especially relevant for histone modifications with narrow enrichment patterns. To provide an example, we analyzed peak coverage of ENCODE HeLa-S3 H3K4me3. Although ChIP-Seq datasets typically have enough resolution to separate H3K4me3 signal upstream of TSSs from the signal downstream (Figure 4.3a), many peak finders can not recover this resolution. Out of the peak finders we tested, only JAMM and PeakRanger can, on average, resolve neighboring H3K4me3 peaks, while other peak finders detect, on average, one large peak encompassing multiple enriched sites (Figure 4.3b and c).

Peak Scoring and Sorting

Instead of attempting to provide a "high-confidence" set of peak calls based on an arbitrary cutoff, JAMM typically reports a large number of peaks and relies on its peak scoring to provide a robust ranking of the reported peaks. This facilitates downstream

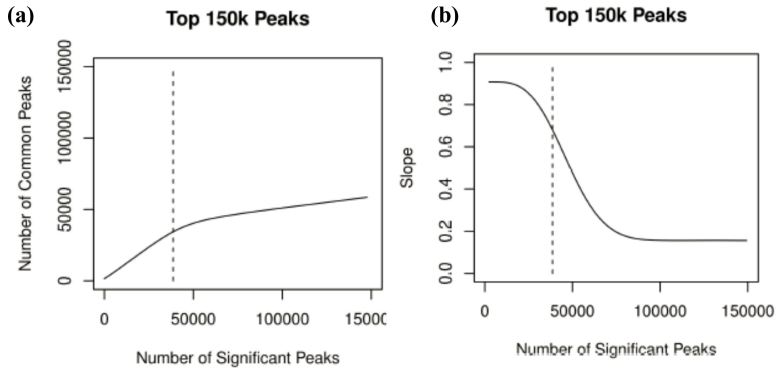


Figure 4.4: Results for IDR analysis on biological replicates for ENCODE HeLa-S3 CTCF using JAMM. Dashed vertical line corresponds to the number of peaks selected with an IDR threshold of 0.02 (38853 peaks). Input to the IDR pipeline included the top 150000 peaks called by JAMM. The number of matched peaks increases as one descends through the sorted peak list up to the point where peaks become irreproducible between replicates. Figure adapted from [73]

analysis and gives users more flexibility in choosing a method to filter the peaks. Irreproducible Discovery Rate (IDR) is an ENCODE-recommended method for filtering peak calls based on replicate reproducibility [331]. The IDR pipeline involves calling peaks on the replicates separately, followed by applying the IDR statistical model to determine the number of reproducible peaks "*n*" given a certain IDR threshold. Peak reproducibility involves whether the peaks overlap and how their ranks compare in the replicates peak lists. Finally, peaks are called on the combined replicates (usually via pooling the aligned reads) and the top *n* peaks are taken to be the high-confidence reproducible peaks. To demonstrate JAMM's peak scoring, we applied the IDR analysis pipeline to HeLa-S3 CTCF ChIP-Seq ENCODE dataset [330]. We found that sorting the peaks using JAMM's peak scores produces a clear phase shift between reproducible peaks and irreproducible peaks (Figure 4.4).

Taken together, JAMM provides a plausible approach to replicate integration that is widely applicable to different types of data. The analysis pipeline would start with peak calling on the replicates separately, followed by IDR analysis to select *n* (the number of reproducible peaks). Finally, peaks are called on the replicates jointly via JAMM's replicate integration and the top scoring *n* peaks are taken as a highly confident set.

4.3.2 High-Resolution Chromatin States[3]

Given the accurate spacial resolution of JAMM, it is feasible to utilize its output in order to obtain high-resolution chromatin states. The main idea is to use JAMM's histone modification peak calls in order to "semi-binarize" the genome. Meaning that instead of using the entire genome-length read count vectors to co-cluster histone modification

[3]Text in this section is largely copied directly from [298]

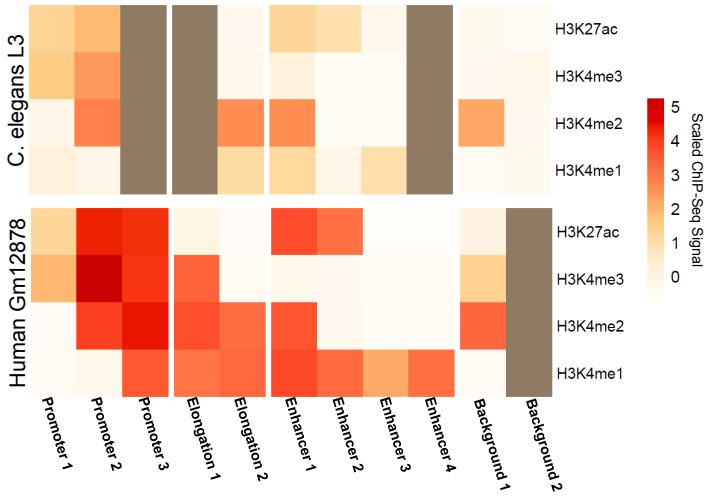


Figure 4.5: Chromatin state definitions based on JAMM+HMM clustering of histone modification ChIP-seq signal at 10-bp resolution for human (gm12878 cell line) and C. elegans (L3 stage). Each state is a multivariate Gaussian distribution. Shown are the distribution mean vectors representing scaled, normalized ChIP-seq signal. Gray boxes indicate the state was not recovered in the respective cell type.

ChIP-Seq data, we use only the signal when there is a peak detected by JAMM and zeros elsewhere. Furthermore, we use an HMM with multivariate Gaussian emissions therefore accounting for differences in signal across genomic regions and for histone modifications covariance.

The chromatin state pipeline starts with defining relevant locations (positions intersecting a ChIP-Seq peak) for each histone modification separately across the genome at 10bp resolution. The signal at relevant locations is defined as background-normalized, extended-read counts. The resulting 10bp binned signal tracks for all histone modifications are matched up and bins that have a zero ChIP-Seq signal in all tracks are discarded. Bins that have a zero ChIP-Seq signal in one or more histone modification track(s) but not the other(s) are assigned a simulated normally-distributed background signal with a mean equal to the lowest bin signal value in the corresponding histone modification track and a variance of 0.1. ChIP-seq signal for each histone modification track is then scaled so that the minimum value is zero and the maximum value is 1000 and converted to log-space.

To learn the emission and transition parameters of the HMM, we employ the Baum-Welch algorithm, initialized via k-means. This learning process results in distinct chromatin states, each represented as a multivariate Gaussian distribution. The mean vector for each state defines the average ChIP-Seq signals of the histone modification tracks in the corresponding state. Finally, we employ the posterior decoding algorithm to

assign a chromatin state to each 10bp bin in the genome that had a peak in at least one of the histone modification tracks. Locations that did not have a peak in any histone modification track (no relevant features, zero signal in all tracks) are not assigned a state. Book-ended bins that have the same state are merged. The output of this process is genome segmentation into variable-width non-overlapping chromatin states similar to Segway [317] and ChromHMM [316].

Figure 4.5 shows the chromatin states learned using this pipeline on H3K4me1/2/3 and H3K27ac in ENCODE [330] GM12878 histone modification data, and modENCODE *C. elegans* L3 histone modification data [193] showing strong concordance between chromatin organization in human and worms, consistent with other similar previous analysis [193].

4.4 Methods

4.4.1 Peak Finding[4]

Data

For ChIP-Seq, we downloaded the fastq files from the ENCODE website [330]. We aligned the reads to the hg19 genome using Bowtie2 with default parameters [59]. We then fltered out reads that did not align uniquely or had more than two mismatches. PCR duplicates were removed using the SAMTools rmdup -s command [332]. Finally, the output from SAMTools was converted to standard BED format using BEDTools [333].

For DNase-Seq, we started with the BAM alignment files from the ENCODE website [330]. PCR duplicates were removed and alignment files converted to BED in the same way as ChIP-Seq.

Comparing Peak Finders

In order any avoid bias due to the number of peaks called when comparing peak finders, we took only the top n ranked peaks from any peak finder, where n is the total number of peaks called by the most strict program (the one that called the least number of peaks). This approach may be biased towards peak finders that have more consistent scores for peak sorting, but this is anyways a desireable property in peak finders.

Motif Analysis

Schmidt *et al.* [334] showed that CTCF recognizes two DNA binding motifs depending on the CTCF domain interacting with DNA. We used the Position Weight Matrices (PWMs) for canonical M1 and M2 motifs identified in Schmidt *et al.* [334] to scan called peaks for motif enrichment.

[4]Text in this section is largely copied directly from [73]

For motif positional enrichment, we extended/truncated called peaks to 750 basepairs in each direction centered at the peak summits (or peak center for peak finders that do not assign a peak summit). We then scanned each extended-peak to obtain a motif log-likelihood score for each position using the SpeakerScan program (parameters: local background window 150 basepairs) [328]. We considered only positive scores and summed the per-position scores across all peaks, and both motifs, to obtain one per-position cumulative motif log-likelihood score for each set of peak calls.

For motif precision, we extended/truncated called peaks to 75 basepairs in each direction centered at the peak summits (or peak center for peak finders that do not assign a peak summit). We then scanned each extended-peak using FIMO motif scanner (parameters: –thresh 0.0001 –max-stored-scores 1000000) [327], using both M1 and M2 motifs separately. Peaks with at least one motif match for either M1 or M2 motifs were counted as positive peak calls.

Peak Coverage Plots

In order to produce peak coverage plots for histone modification peak calls, we intersected each set of peak calls with annotated promoter regions from UCSC hg19 known genes using BEDTools *intersect* command [333], where a promoter region is defined as 4000 basepairs in each direction centered at the annotated TSS. For each set of peak calls, each position in the 8000 basepair window was assigned a score of 1 for each intersecting peak. Per-position scores were summed to produce peak coverage and each position was divided by the mean per-position score to normalize the coverage scores.

Heat maps

To produce read coverage heat maps, we obtained peak regions (1000 basepairs in each direction centered at peak center) and produced smoothed extended-read counts for those regions at 10 basepair resolution, using a slightly modified version of the code developed for JAMM (see Supplementary Methods). For ChIP-Seq, the fragment length predicted by JAMM was used to extend the reads, for DNase-Seq the 5?-ends of the reads were counted. The read counts for every peak region were scaled to be between 0 and 100. Peak edges were also assigned at 10 basepair resolution and overlaid on the read coverage heat maps.

Genome Browser Plots

Extended, RPKM-normalized read counts were produced using the *bamCoverage* command from DeepTools [335]. For histone modifications, reads were extended to 148 basepairs. For CTCF ChIP-Seq, reads were extended to 100 basepairs. The produced bigWig coverage tracks and narrowPeak/bed files for called peaks were visualized in IGV browser [336].

4.4.2 Chromatin State Pipeline

Chromatin state pipeline was performed as described in [337] (see Chapter 5) with the following differences: peaks were called using JAMMv1.0.7rev5 (parameters: -r window -b 150), the signal obtained was not smoothed and the Baum-Welch algorithm was implemented fixing the transition probability matrix to 0.9 at the diagonal and $0.1/(n-1)$ elsewhere where n is the number of states and the segmentation was performed using the posterior decoding algorithm instead of the Viterbi algorithm. The implementations from Baum-Welch and posterior decoding used are available at `https://github.com/mahmoudibrahim/hmmForChromatin`.

4.5 Discussion[5]

Peak Finding

A desirable property in peak finders is the ability to detect, and correctly determine the widths of, enrichment sites with different widths and SNR. Out of the peak finders we tested, we found that only JAMM is able to accurately determine widths of enrichment sites that have different properties and are in the same dataset. This is because JAMM learns the parameters of the Gaussian mixture model for every enriched window independently and only fixes the structure of the covariance matrix (see Pipeline and Supplementary Text). On the other hand, some peak finders start with learning an expected peak shape [70, 72], which makes it more difficult to detect enrichment sites with different widths and assign their boundaries accurately.

Some peak finders adapt various sub-routines for refining peak widths after peak finding is complete (see [66] for example). Depending on the sub-routine used, this approach may be able to assign accurate peak boundaries, but only when the original peak represents one enrichment site. When the original peak represents several enrichment sites that are closely spaced, this approach typically results in picking one of the sites and missing the others (see Figure S3 in [66] for an example). We showed that JAMM's local signal clustering also avoids this caveat and can correctly resolve neighbouring punctate sites, similar to programs specifically designed for this like PeakRanger [312].

We introduced JAMM as a universal peak finder and showed that it can analyze different types of datasets with very little change to the underlying pipeline. This demonstrates that finding enriched sites, in read-density based NGS datasets, is essentially the same task regardless of the sites signal properties. Therefore, developing specialized algorithms is rather unnecessary [72]. Instead, we propose that more attention could be directed towards developing universal peak finding solutions, refining pre-processing of read counts to correct for mappability and structural variation biases [24], and towards developing solutions for biological replicates integration.

There is no consensus on how to analyze biological replicates, although pooling replicates is part of the ENCODE consortium ChIP-Seq recommended pipeline [338].

[5]Text in the "Peak Finding" subsection is largely copied directly from [73]

But when peaks are called on pooled replicates, read counts at every basepair are normalized to the total depth of the pooled reads. Therefore, the differential intensities and differential spatial coverage of the replicates are obscured. We have demonstrated that this differential information is important. JAMM attempts to address replicates integration by considering the covariance of the replicates read counts. To our knowledge, there are no other published peak finders that attempt to integrate biological replicates in order to address spatial resolution and sorting of peaks. We showed that JAMM replicates integration results in improved peak sorting and spatial resolution over peak calling on the pooled replicates. It also appears that JAMM with replicate integration consistently performs better than JAMM with pooled replicates when measured via genome-wide motif content benchmarks. Nevertheless, it should be noted that motif-content benchmarks do not represent a definite "gold standard" due to our incomplete understanding of protein-DNA interactions and potential biases in the methods used to produce the benchmarks.

In fact, despite the lack of a systematic benchmark for peak finidng [339], many researchers focus on improving specificity and use specificity-related benchmarks to show that a program outperforms the others (see [66, 72] for examples). In addition, some peak finders can not adjust to reporting a larger number of peaks and/or ignore providing appropriate peak scores. However, due to the experimental and biological variability and the typically large number of cells involved in NGS protocols, one can not rule out that a peak with less confidence may be more relevant in the down-stream analysis than a peak with higher confidence. Therefore, a more sound approach would be to report a large number of peaks but with appropriate peak scores [307, 331]. Appropriate peak scores would have few or no ties and represent the confidence in the peak accurately based on its read density and how it compares to control. A user can then either use the full peak list and combine it with other evidence, or choose the top scoring peaks via empirical False Discovery Rate analysis [340] or Irreproducible Discovery Rate (IDR) analysis [331]. Owing to its relaxed enriched window determination routine, JAMM typically determines a large number of peaks. But JAMM also provides a robust peak scoring system with very few score ties, if any. This, in addition to its accurate peak width determination, makes JAMM suitable for the IDR pipeline; which is recommended by the ENCODE consortium to determine biological replicates reproducibility in ChIP-Seq datasets [338].

Areas where JAMM can be improved include its background model (when there is no biological control available); which does not take into account mappability and structural variations. Biases due to mappability and structural variations are especially relevant for cancer cell lines for example [24, 25, 66]. In the near future, we might incorporate structural variation correction in a manner similar to that implemented in F-seq [64]. Finally, for some of the ChIP-Seq datasets published, the sample SNR will be less than that of the biological control; a problem that has been reported earlier [65]. JAMM, as well as other peak finders, fail to analyze those samples properly because of the normalization to the total depth of the sample. One way to remedy this would be to separate reads into signal and background prior to normalization [341].

Chromatin States

Combining multiple histone modification features into one integrated model in a completely unsupervised manner is a challenging task because of the complex noise structure in the data, the over dispersion of the data and the complex dependencies between the different features. Another possible challenge is the resolution to which one desires to summarize the data which can greatly affect the interpretation of the results. Summarizing the data at a very coarse resolution can help identify major types of genome regulation [314, 342] while more resolved approaches like ChromHMM [315, 316] help discover different types of regulatory elements. Here we segment the genome at 10-basepair resolution enabling the discovery of the spatial organization of histone modification combinations around regulatory elements (see Chapter 5).

"The main advantage of our chromatin state genome segmentation pipeline is that it allows for chromatin state assignment at high-resolution using semi-binarized signal, as opposed to using fully binarized (enriched / not-enriched) information at 200 bp resolution utilized in the ChromHMM approach [315, 316]. Our semi-binarized signal is the extended ChIP-Seq read counts for relevant locations in the genome (ChIP-Seq peaks) and zeros elsewhere. Therefore, information about the co-variance of the histone modifications' signals can be included, but without suffering from noise overrepresentation. This has the potential to lead to more meaningful clustering of the histone modification signals compared to previous approaches [317]. Finally, we do not analyze the entire genome, but only locations which had ChIP-Seq peaks in at least one histone modification dataset. Therefore, we can assign chromatin states at high-resolution 10 bp bins, close to the single-basepair resolution of Segway [317] but without its expensive computational resources requirement"[298]. Possible improvements on the chromatin state pipeline described here is to model read counts directly using count data probability distributions instead of the Gaussian distribution (similar to the approach followed in [321]) and to include a Dirichlet prior which can help automatically determine a valid number of chromatin states.

Chapter 5

Promoter Chromatin Directionality

5.1 Contribution Statement

Results in this chapter are based on work in the following publication: [298].

Contributions to Chapter Results: Sascha Duttke (Kadonaga Lab / UC San Diego) and Sven Heinz (Glass Lab / UC San Diego) produced the 5'GRO-Cap data. Scott A Lacadie (Ohler Lab / Max-Delbrueck-Center, Berlin) analyzed the 5'GRO-Cap data and performed the open regions stratification based on 5' GRO-Cap data. Mahmoud M Ibrahim analyzed the DNase-Seq data and histone modification ChIP-Seq data and performed the chromatin state segmentation and all its associated analyses.

5.2 Introduction

Bidirectional transcription, in which two genes are arranged in a head-to-head fashion upstream antisense to each other (Figure 5.1), has been noted as an important regulatory feature as early as the 1980s [343] and is hypothesized to allow certain genes to be co-regulated in a multitude of ways [344–346]. Bidirectional transcription initiation typically occurs from the same promoter nucleosome-free region (NFR) leading to the hypothesis that it could have a role in stabilizing nucleosome positions thereby stabilizing gene activation response [347].

Divergent transcription is the term applied to bidirectional transcription when one transcript is stable but the other is unstable and rapidly degraded. Divergent transcription was observed to be pervasive in yeast [348, 349]. Intriguingly, this was observed to be also a common feature in mouse [350], human [51] and *C. elegans* [56, 351] but not in *D. melanogaster* [352]. To answer the question of why this feature evolved in some eukaryotes but not others, one must first ask how these divergent transcripts are initiated and regulated.

Transcription start sites of stable and divergent transcripts were shown to coincide with two separate transcription preinitiation complexes (PIC) in yeast using high-resolution ChIP-Seq (ChIP-exo) [353]. Additionally, recent work in the Kadonaga and

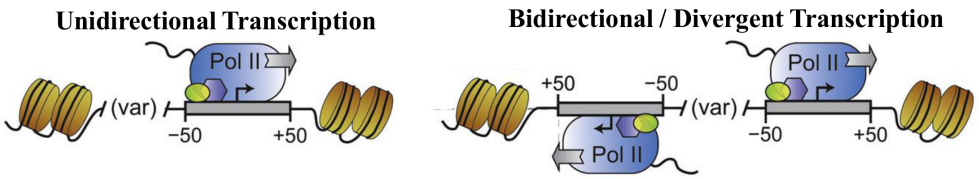


Figure 5.1: Models of Transcription Initiation Directionality. Figure adapted from [298]

Ohler labs showed, both *in vivo* and *in vitro*, that divergent transcription in human occurs from two separate core promoters at the edges of promoter NFRs and that each core promoter is by itself unidirectional [298, 337], which is in agreement with other concurrent work [354, 355].

Hence, a consensus model arises in which divergent transcription is initiated via two separate PIC complexes from two separate core promoters oriented in an upstream antisense direction to each other at the edges of promoter NFRs (Figure 5.1). In this model, it is also conceivable that one might observe unidirectional promoter NFRs if there is only one core promoter in the promoter NFR (Figure 5.1). Such cases were characterized in [298] and [355]. In fact, even in divergent promoter NFRs, the directionality of transcription initiation (defined as the ratio of the initiation rate of the sense core promoter to that of the antisense core promoter) can have a wide range depending on factors such as core promoter strength and the promoter chromatin environment [8, 337]. But how does promoter chromatin relate to divergent transcription?

The availability of genome-wide assays for histone modifications, DNA accessibility and nascent transcription facilitates the comprehensive investigation of promoter chromatin environment in the light of pervasive divergent transcription. Since promoter NFRs are typically in the order of 100-200bp wide with core promoter sequences lying at their edges, this task involves high resolution analysis of genome-wide sequencing DNase-Seq data in order to accurately demarcate the edges of open chromatin, as well as high-resolution analysis of ChIP-Seq data in order to resolve promoter histone modification combinations at high resolution.

5.3 Results

We started by applying a modified version of the chromatin state pipeline explained in Chapter 4 (see Methods) to HeLa-S3 H3K4me1/2/3 and H3K27ac histone modification data from HeLa ENCODE ChIP-Seq data [330] and obtained 8 chromatin states typical of promoter and enhancer chromatin (Figure 5.2). We then defined open regions with accurate widths from HeLa ENCODE DNase-seq data [330] using JAMM (see Chapter 4, [73]) and stratified those open regions, using gene annotations and 5'GRO-Seq data (see [298] / Sascha Duttke and Scott A Lacadie), into divergent open regions (open regions with two initiation only one of which initiates a stable gene, Figure

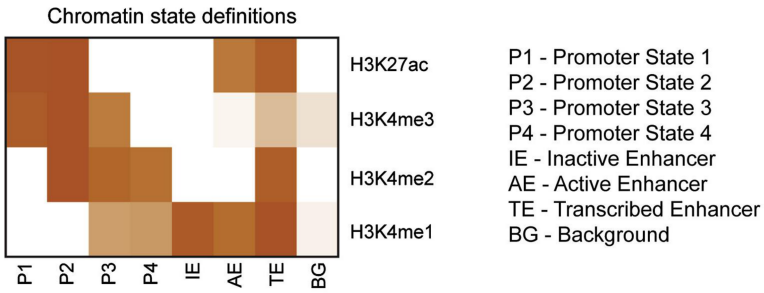


Figure 5.2: Chromatin state definitions based on JAMM+HMM clustering of histone modification ChIP-seq signal at 10-bp resolution. Each state is a multivariate Gaussian distribution. Shown are the distribution mean vectors representing scaled, normalized ChIP-seq signal. Figure adapted from [298]

5.2), unidirectional open regions (those with only initiation event in the stable gene direction, Figure 5.2), bidirectional open regions (those with two initiation events both of which initiate stable genes, Figure 5.2) and enhancer RNA [289, 290] open regions (open regions with two initiation events both of which annotate to intergenic locations, enhancers are discussed in Chapter 3).

Using those four open region groups and the JAMM+HMM chromatin states, we made two observations:

(1) We plotted the chromatin state coverage at the four classes of open regions (Figure 5.3). "We found that in the stable gene direction, one observes a clear cascade of chromatin states where H3K4me3 and H3K27ac are found together at the +1 nucleosome location ("promoter state1"), followed by the gain of H3K4me2 ("promoter state 2"), then the loss of H3K27ac ("promoter state 3"), and finally the loss of H3K4me3 ("promoter state 4"). However, in the unstable transcript direction in divergent open regions, one observes an enrichment of "promoter state 2" immediately downstream of the TSS (on the -1 nucleosome) and no enrichment of "promoter state 1". Furthermore, there is no preference for any particular chromatin state on the reverse side of unidirectional open regions" [298]. Finally, eRNA open regions feature an altogether different cascade, with three different enhancer-related chromatin states. Therefore, divergent promoter NFRs feature a unique and directional combination of histone modifications distinct from that found in unidirectional promoter NFRs, bidirectional promoter NFRs and intergenic enhancer RNA NFRs.

(2) Given the high-resolution of our analysis and the observed directionality of divergent open regions, we can focus the question further: are histone modifications downstream of the unstable transcript in divergent promoter open regions correlated with the forward transcript initiation rate and vice versa? We counted H3K27ac reads in the 148bp regions downstream and upstream of divergent promoter NFRs and correlated H3K27ac read counts with 5'GRO-Seq counts as a measure of initiation rate (Table 5.3). We observed that the stable gene initiation rate correlates with H3K27ac levels downstream of the TSS but not with the H3K27ac levels upstream of the open

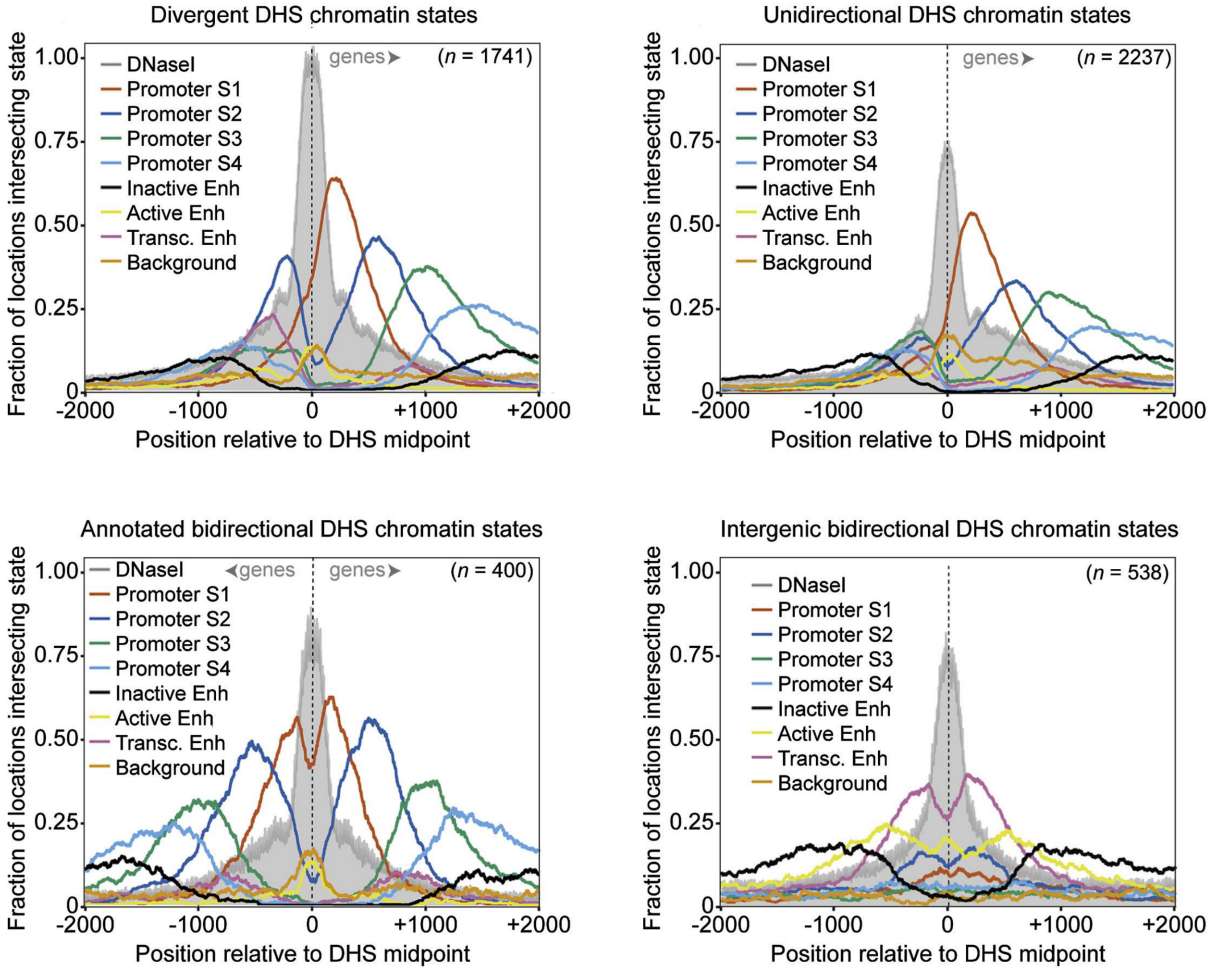


Figure 5.3: Chromatin state coverage 2 kb around the center of divergent promoter NFR, unidirectional promoter NFR, bidirectional promoter DHS NFR, and divergent intergenic NFR at single nucleotide resolution. Grey = DNaseI-seq read 5 end counts, red = Promoter State1, blue = Promoter State 2, green = Promoter State 3, light blue = Promoter State 4, black = Inactive Enhancer, yellow = Active Enhancer, pink = Transcribed Enhancer, and orange = Background. See Figure 5.2. Figure adapted from [298]

region (that is, downstream of the unstable transcript TSS). In addition, the initiation rate of the unstable transcript correlates with the H3K27ac downstream of the unstable TSS but not with that downstream of the stable gene TSS. Therefore, histone modification levels are also directional in the sense that they are linked to the initiation rate of their corresponding TSS and not the TSS on the other side of the promoter NFR.

	Initiation Rate - Forward	**Initiation Rate - Reverse**
H3K27ac - Forward	0.39 (p <0.0001)	0.0001 (p = 0.09)
H3K27ac - Reverse	0.04 (p = 0.09)	0.25 (p <0.0001)

Table 5.1: Spearman Rho correlation values are shown with corresponding p values between transcription initiation rate (measured using 5'GRO-seq read 5end counts) and H3K27ac histone modification (measured using ChIP-seq fragment-extended read counts intersecting a window 148 bp downstream of the appropriate DHS peak edge). Table adapted from [298].

5.4 Methods[1]

5'-GRO-seq data was produced as described in [298] (Sascha Duttke / Kadonaga Lab).

All 5 datasets of ENCODE-mapped DNase-seq reads for HeLa-S3 cells were downloaded from the UCSC ENCODE ftp server [330]. PCR duplicates from each file were removed using SAMTools [332]. The resulting files were converted to BED using BEDTools [333] and concatenated before peak calling with JAMM v1.0.6 [73] (settings: -m narrow -f 1). HeLa-S3 cell, Broad Institute histone modification ChIP-seq raw fastq files were downloaded from the UCSC ENCODE ftp server [330]. Reads were aligned to hg19 genome using Bowtie2 [59] with default parameters and then filtered for those that did not align uniquely or had more than two mismatches. PCR duplicates were removed after alignment using SAMTools [332] and converted to standard BED format using BEDTools [333]. Histone modification peaks were called using JAMM v1.0.4rev1 [330] with default settings while maintaining all replicates separate. The filtered peak lists produced by JAMM were considered for further analysis.

In order to define promoter NFRs as divergent or unidirectional, BEDTools [333] intersect command was used to find overlaps between DNaseI-seq peak calls (defining NFRs) and 5-GRO-seq cluster modes, both described above. The output from BEDTools was then parsed with custom Perl scripts into different NFR categories. NFRs with exactly one intersecting TSS cluster mode were considered unidirectional. NFR with exactly two intersecting 5-GRO-seq cluster modes where the two modes were upstream and antisense of each other, one annotating as TSS and the other as intergenic, were considered divergent. NFRs with more than one intersecting 5-GRO-seq cluster modes on any one DNA strand, or with two 5-GRO-seq cluster modes on opposite

[1] Text in this section is largely copied directly from [298]

strands but downstream of each other, were removed from further analysis. For an increased-confidence unidirectional group, unidirectional classified NFRs intersecting reverse-side annotated TSSs (yet having no 5GRO-seq clusters) or containing exactly one TSS-annotating cluster mode that was also part of the divergent or bidirectional reciprocal closest upstream antisense selection (described above) were considered ambiguous and removed from further analysis.

We employed a Hidden Markov Model (HMM) for unsupervised genome-wide clustering of histone modification ChIP-Seq read counts (also see Chapter 4). We chose a multivariate Gaussian distribution for the HMM state emissions. Each chromatin state is a multivariate Gaussian distribution fully defined by its means vector, corresponding to the signals' means of the histone modification tracks, and its covariance matrix.

In a pre-processing step, we define relevant locations for each histone modification (positions intersecting a ChIP-Seq peak) separately across the whole genome at 10-basepair resolution. The signal at relevant locations is defined as background-normalized, smoothed, extended-read counts (ie. ChIP-Seq signal). Peaks were identified using JAMM [73], as described above. For each histone modification dataset, we extracted the corresponding ChIPSeq signal for each peak at single-basepair resolution, using the SignalGenerator pipeline provided with JAMM [73]. JAMM's SignalGenerator output is then aligned to the genome in 10-basepair bins using the BEDOps [356] bedmap command (settings: –mean). Bins that did not intersect ChIP-Seq peaks are assigned a signal of zero. ChIPSeq signal for each histone modification track is then scaled so that the minimum value is zero and the maximum value is 1000 and converted to log-space.

The resulting 10-basepair binned signal tracks for all histone modifications are matched up and bins that have a zero ChIP-Seq signal in all tracks are discarded. Bins that have a zero ChIP-Seq signal in one or more histone modification track(s) but not the other(s) are assigned a simulated normally-distributed background signal with a mean equal to the lowest bin signal value in the corresponding histone modification track and a variance of 0.1. To learn the emission and transition parameters of the HMM, we employ the Baum-Welch algorithm as implemented in the RHMM R package, initialized via k-means, on the signal tracks of chromosome 1. This learning process results in distinct chromatin states, each represented as a multivariate Gaussian distribution. The mean vector for each state defines the average ChIP-Seq signals of the histone modification tracks in the corresponding state. We 0-to-1 scale the means across each histone modification to define the prototypical chromatin states shown in Figure 5.2.

Finally, we employ the Viterbi decoding algorithm [22] as implemented in the RHMM R package to assign a chromatin state to each 10-basepair bin in the genome that had a peak in at least one of the histone modification tracks. Locations that did not have a peak in any histone modification track (no relevant features, zero signal in all tracks) are not assigned a state. Book-ended bins that have the same state are merged. The output of this process is genome segmentation into variable-width non-overlapping chromatin states similar to Segway [317] and ChromHMM [316]. To

produce chromatin state coverage plots, we started with windows defined around the midpoints of NFRs as described above. Chromatin states were intersected with NFR-based windows using BEDTools [333] intersect command.

5.5 Discussion

Transcription initiation is regulated by a variety of sequence and chromatin signals that define transcription start sites and promoter regions. Using genome-wide high-throughput sequencing data, it is possible to investigate general mechanisms of promoter regulation. Using JAMM-defined open regions and JAMM+HMM chromatin states (see Chapter 4), we were able to establish that promoter chromatin architecture in human is directional and intimately linked to the transcription initiation rate and the directionality of promoters [298, 337]. More formally, we found that (1) divergent promoter NFRs in HeLa-S3 and other human cell lines (not shown) feature a unique histone modification combination (ie. chromatin state) not enriched in other regulatory regions with transcription initiation and that (2) histone modification levels correlate to transcription initiation rate in a directionally-specific manner.

This result poses various questions. For example, why are unstable divergent transcripts marked by a different chromatin state than that marking stable transcripts? One hypothesis explored in [8] and [354] is that the level of transcription initiation is related to the level of methylation on the closest downstream nucleosome. Since divergent transcripts are on average lower expressed than sense stable transcripts [298], it is expected to find exclusively H3K4me3 on both H3 copies of the sense nucleosome and a "reduced" methylation level on the antisense nucleosome with asymmetric H3K4me3/H3K4me2. Of course "it is important to note that due to the two copy nature of histones within nucleosomes, it is not possible to distinguish single, symmetric, or asymmetric modifications of a single histone octamer using typical cell population ChIP-based chromatin states. Colocalized [histone modification] signals could reflect mixtures of different cell populations in the biological sample, differences between alleles, indirect cross-linking of distant loci colocalized in three-dimensional space, or the low resolution of standard ChIP-seq, rather than the physical presence of multiple marks within a single nucleosome" [8].

A slightly different question is whether histone modification *levels* and the histone modification *combinations* (ie. chromatin states) warrant separate consideration in the analysis of chromatin information, given the nature of cell population-based assays. The results described here indicate that while histone modification levels are related to the percentage of cells that have a certain modification in a given location, the level of transcription or the stability of the nucleosome in that location, chromatin states could indicate the regulation "mode" of a particular regulatory element regardless of what the actual histone modification levels are.

Another issue concerns predicting expression from histone modification data. Previous studies have shown that histone modification levels are predictive of gene expression and of each other's levels ([218, 219, 318], see Chapter 3), "where it is

common practice to consider signal within large windows around annotated, or experimentally measured" [8] promoter regions. This work provides an example where a higher-resolution analysis might be useful in terms of delineating the directionality of chromatin states in promoters. In fact, the detailed spatial arrangement of histone modifications around regulatory elements is often overlooked perhaps due to the perceived resolution limits of ChIP-Seq. We have shown that given high quality ChIP-Seq data and careful highly resolved models for analysis and integration, interesting patterns regarding the spatial arrangement of histone modifications at regulatory elements can be observed.

Finally, it is worth noting that we explain the results in terms of the predominant chromatin state at a particular location relative to regulatory elements (for example, "promoters feature promoter state 1 immediately downstream of the transcription start site"). By this we intend to highlight the predominant average trend, meaning that not all stable transcripts will feature the same chromatin state. Other promoters featuring other states could indicate lower transcription level (see above), a different mode of transcription initiation regulation, noise in the data or errors in the classification by the models and algorithms used. This and all the other questions discussed here can potentially be resolved with more careful higher resolution ChIP experimental technologies (see Chapter 2).

Chapter 6

Promoter Dynamics during Motor Neuron Programming

6.1 Contribution Statement

Results in this chapter are based on work in the following publication: [155].

Contributions to Chapter Results: Silvia Velasco (Mazzoni Lab / New York University) produced all the ChIP-Seq data. Mohamed Al-Sayegh (Mazzoni Lab / New York University) produced all the RNA-Seq data. Mahmoud M Ibrahim developed the Bayesian Network clustering model, performed all the ChIP-Seq, RNA-Seq and associated chromatin state analyses.

6.2 Introduction

Promoter chromatin dynamics during stepwise *in vitro* differentiation have been profiled using time-course ChIP-Seq in a few select differentiation systems (see Chapter 3). However, it is not clear how promoters behave in direct programming or transdifferentiation protocols. In the rapid and efficient NIL differentiation system, cell fate is completely changed within only 48 hours [155, 160] (see Chapter 3). Therefore, promoter chromatin remodeling, activation and repression have to occur within a short span of time under dynamic constraints that are otherwise alien to the gene regulation machinery, since the same cell fate transformation in *in vivo* development takes weeks.

In collaboration with the Mazzoni lab, we take advantage of the NIL system and profile promoter-related histone modifications at multiple time-points during NIL-induced programming of motor neurons using ChIP-Seq ([155], Silvia Velasco / Mazzoni lab). I will describe a Bayesian Network model for clustering of multiple histone modification data across time (see Chapter 2 for an introduction to time-course clustering) and apply it to this time-course histone modification data, in order to ask: (1) do all motor neuron genes exhibit the same activation dynamics?, (2) do all pluripotency genes exhibit the same repression dynamics?, (3) is bivalency required for promoter activation during differentiation? and (4) what are the dynamics of acetylation

and methylation of H3K27 during repression and activation of promoters? Answering those questions is important in terms of delineating promoter chromatin dynamics during differentiation and also in the context of understanding the NIL differentiation system which is an efficient and homogeneous *in vitro* differentiation protocol [155, 160] (see Chapter 3).

6.3 Results

6.3.1 Clustering Combinatorial Time-course Data using Bayesian Networks with Tree-like Structures[1]

Two key ideas are important to the model described here for co-clustering of multiple genome-wide data sets over time. First, cell differentiation often follows hierarchical trajectories which, from a graph theory point of view, exhibit a tree structure. Therefore, when building a Bayesian network to model such processes, a tree-structured network is a natural representation. Second, when investigating time-course changes in genome-wide data, we are often interested in changes occurring over time and not necessarily in absolute levels of histone modifications (see above). Therefore, directly clustering fold-change values (the logarithm of the ratio of one time point relative to the other) is a simple way of implementing this idea and it is especially convenient because it resolves many issues in terms of normalization and comparison of the data, as opposed to attempting to cluster ChIP-Seq read counts. Additionally, log-fold change values typically follow a Gaussian distribution which is convenient for implementation and design of the Bayesian Network model.

Given those two ideas, we designed a Bayesian network with a conditional Gaussian probability distribution (Figure 6.1) [357]. The model features one discrete unobserved class variable C_i upon which all continuous univariate Gaussian observed variables are conditioned. The discrete unobserved variable represents the cluster that defines a certain chromatin state trajectory, while the continuous observed variables represent the consecutive log2 fold-changes in ChIP-seq signal between the consecutive time points. To ensure sparsity and avoid large covariance matrices, each histone modification is modeled via its own tree, meaning that each histone modification is independent of all other histone modifications given the discrete class variable. This gives a structure similar to a Naive Bayes model (see Chapter 2) in terms of independence between different chromatin marks, but we allow for dependencies between the observed variables representing a histone modification at different time points as long as the acyclicity condition of BNs is satisfied (Figure 6.1). Each univariate Gaussian node is modeled via linear regression of its corresponding univariate Gaussian parent [357]. Since any continuous node will also be conditioned on the unobserved discrete class node, a different set of regression parameters is defined separately for each value of the discrete parent (ie. each cluster defines a different chromatin state trajectory).

[1] Text and figures in this section are largely copied directly from [155]

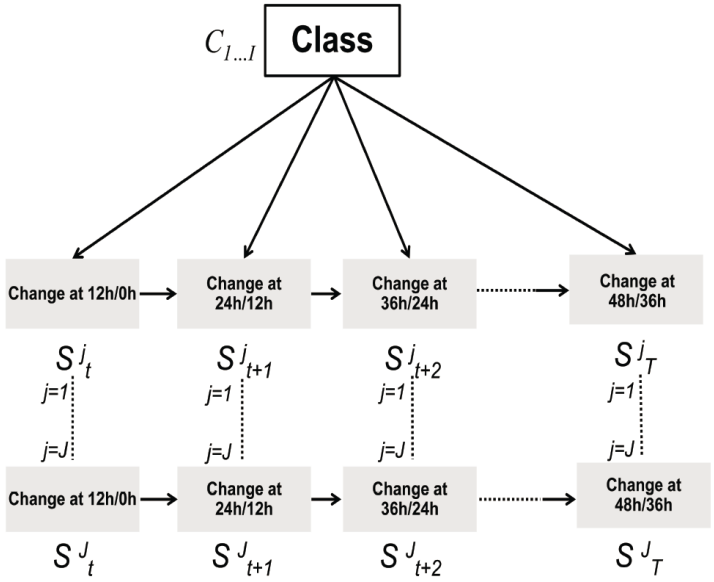


Figure 6.1: A graphical representation of the conditional Gaussian Bayesian Network for co-clustering of multiple time-course histone modification data. Figure adapted from [155].

If only one histone modification is modeled, the network reduces to a Tree-augmented Naive Bayes model [358].

This model can be summarized as follows:

$$p(C_i, S_{t=1}^{j=1}, ..., S_T^J) = p(C_i) \prod_{j=1}^{J} \left[\prod_{t=1}^{T} p(s_t^j | c_i, s_{t-1}^j) \right]$$

where $C_{i=1}^I$ denotes the class discrete variable with space $i = 1, ..., I$ and s_t^j denotes a univariate Gaussian distribution where $t = 1, ..., T$ are the T time-points modeled and $j = 1, ..., J$ are J histone modifications. Note that T is equal to $N_d - 1$ where N_d is the number of time points assayed, since we model the fold-change values and not the actual count data. Each Gaussian node is a univariate Gaussian distribution whose mean μ is a linear function of its continuous parent if any:

$$s_t^j \sim N\left(\alpha + \beta \mu_{s_{t-1}^j}, \sigma^2_{s_t^j}\right)$$

where $\alpha(c_i)$ and $\beta(c_i)$ are positive real numbers. However, the regression parameters α and β, and the variance of the Gaussian node $\sigma^2_{s_t^j}$ are also conditional on the discrete unobserved node C. Therefore, the conditional Gaussian distribution $\mathcal{L}$ of any $(s_t^j | c_i, s_{t-1}^j)$ is given by

$$\mathcal{L}(s_t^j|c_i, s_{t-1}^j) = N\Big(\alpha(c_i) + \beta(c_i)s_{t-1}^j, \sigma^2\big(s_t^j(c_i)\big)\Big)$$

where $\sigma^2(s_t^j)$ is the variance of s_t^j. When $t = 1$, any $s_{t=1}^j$ will have no Gaussian parents and will be conditional only on the discrete class variable C. In that case, $\mathcal{L}(s_{t=1}^j|c_i)$ is given by:

$$\mathcal{L}(s_{t=1}^j|c_i) = N\Big(\mu(c_i), \sigma^2\big(s_{t=1}^j(c_i)\big)\Big)$$

where $\mu(s_{t=1}^j)$ is the mean of $s_{t=1}^j$ and $\sigma^2(s_{t=1}^j)$ is the variance.

6.3.2 Promoter Dynamics during Motor Neuron Programming[2]

In order to understand promoter chromatin dynamics during neuron differentiation, we obtained ChIP-Seq data for H3K4me3, H3K27me3 and H3K27ac at four time-points during NIL-induced motor neuron programming of ESCs ([155], Silvia Velasco / Mazzoni lab). We defined promoter regions as -200bp to +2000bp at all annotated Gencode mm10 TSSs (vM3) [359], in order to avoid up-stream histone modifications probably related to divergent transcription (see Chapter 5).

"We applied the Bayesian network model described above to cluster promoter regions based on the combinatorial trajectories of H3K4me3, H3K27ac and H3K27me3 histone modifications into 11 promoter classes (P1 to P11). To choose the number of clusters, we used 10-fold cross-validation and examined the change in the likelihood of the model as the number of clusters increases (Figure 6.2). Although we did not find evidence of over fitting for the range examined (up to 20 clusters), cluster numbers higher than 11 improve the model likelihood only modestly. [...] We chose 11 clusters as a good balance between ease of interpretation and the fit of the model to the data (Figure 6.2). Grouping those promoter classes into three broad groups for upregulation, downregulation and no-change reveals that promoters follow multiple distinct activation and repression trajectories, which in turn correspond to distinct gene expression dynamics (Figure 6.3). This is reflected in the extent of up- or downregulation as well as the slope of change in gene expression. Scaling the expression of each gene and visualizing the scaled values as a heat map shows that different promoter groups correspond to different up- and down-regulation kinetics (Figure 6.4)." [155]

"The highest promoter and transcription activation occurs in P1 promoters, which start in a bivalent H3K4me3/H3K27me3 state and resolve into an active H3K4me3/H3K27ac state (see Chapter 3). Gene Ontology (GO) and Reactome pathway enrichment analysis shows that those genes are enriched in motor neuron differentiation and axonogenesis genes (not shown). In contrast, P7 promoters show an opposite trend where they start in an active H3K4me3/H3K27ac state and switch to a repressed H3K27me3 state, also reflected in a strong and rapid decrease in gene expression. GO and Reactome

[2]Text and figures in this section are largely copied directly from [155]

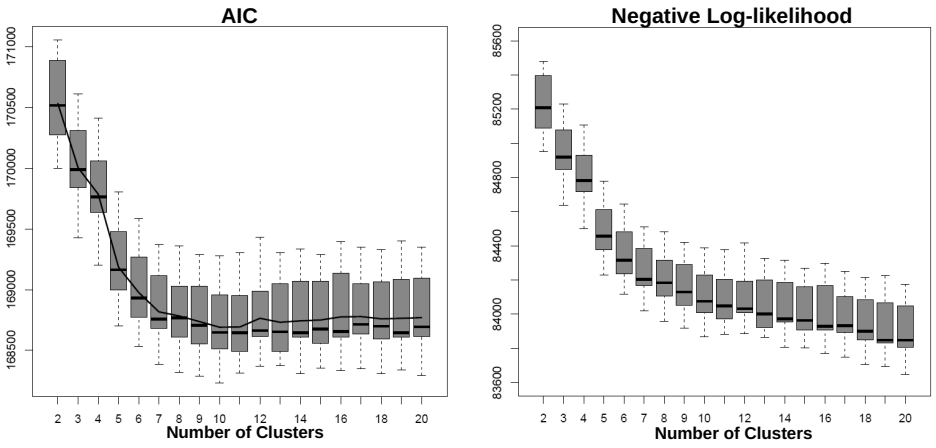


Figure 6.2: 10-Fold cross validation of the Bayesian Network model to choose the number of promoter classes. As the number of promoter classes increases the model better explains the data and the model likelihood given the data improves. Choosing 11 promoter classes strikes a balance between model fit to the data and increased model complexity. Right: Negative Log-likelihood, Left: Akaike Information Criterion (line represents airthmetic mean). Figure adapted from [155].

analysis show enrichment for pluripotency genes in this group (not shown). Similar to P1 promoters, P10 promoters start in a bivalent H3K4me3/H3K27me3 state, but are not activated during differentiation. GO analysis indicates a general enrichment for cell fate specification showing that this group includes cell-fate specific genes that are not activated during motor neuron differentiation (not shown). The contrast between P1, P7 and P10 promoters suggests that during NIL induction pluripotency genes (e.g. Lin28a, Fgf4, Oct4 and Sox2) are repressed as stem cell fate is extinguished presumably by the activity of the programming factors and culture conditions, while neuron (e.g.Tubb3) and motor neuron genes (e.g. Chat, Isl2 and Hb9) are activated and genes related to other developmental pathways are unchanged (e.g. Tead4, Tbx5, GATA6)" [155].

"Therefore, NIL induction in a chromatin environment distinct to that encountered during normal development results in significant promoter chromatin remodeling consistent with a motor neuron fate. Further, these results reveal that even without transitioning through progenitor stages, bivalent chromatin states at promoters get resolved in a lineage specific manner as they do during stepwise differentiation" [155]. Of note, promoters starting in a bivalent state seem to become active at a faster rate than promoters starting at silent (P2), ambiguous (P3) and active (P4) states (Figures 6.3 and 6.4), consistent with previous results indicating that repressed chromatin has faster response to activation signal than already active chromatin [360]. Therefore, it is likely

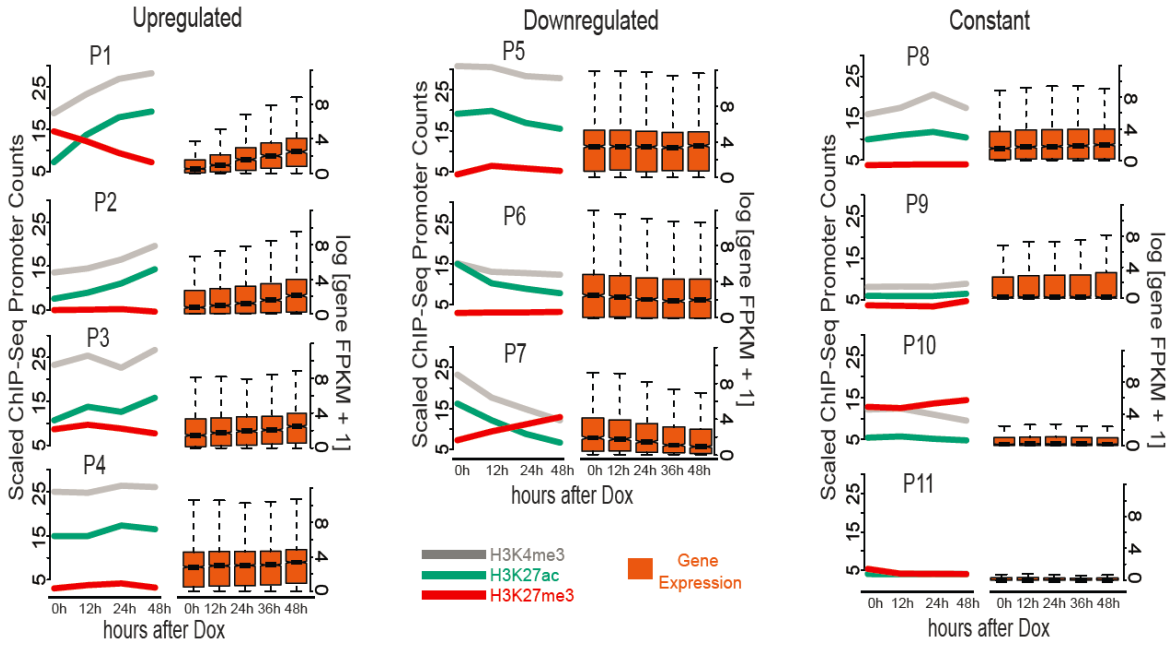


Figure 6.3: Promoter classes based on combinatorial histone modification dynamics at promoters classified using a Bayesian Network model for time-course chromatin states (left) and their corresponding gene expression levels (right). Distinct dynamics of promoter activation/inactivation are related to distinct dynamics of gene expression. ChIP-Seq values displayed are averaged for each promoter region and linearly scaled to ensure different histone modifications are comparable. Figure adapted from [155].

that bivalent chromatin state is not required for subsequent activation. Indeed, classifying promoter regions by whether they are in a bivalent chromatin state at 0h reveals that only P1 and P10 start in a clear bivalent state (Figure 6.5), also consistent with previous results showing that abolishing bivalency by inhibiting the deposition of H3K4me3 in otherwise bivalent promoters does not abolish activation during differentiation [248].

6.3.3 Bivalent Promoters Transition through a Trivalent Chromatin State

Examining the trajectories of H3K27me3/ac in promoter regions belonging to P1 and P7 groups show an acetylation / methylation switch at H3K27 in the period from 12h to 24h, suggesting that bivalent chromatin state is resolved by transitioning through a trivalent H3K4me3 / H3K27me3 / H3K27ac state. To further refine this observation, we built a static chromatin state model (see Methods and Chapters 4 and 5) for promoter regions using the four histone modifications H3K4me3/2 and H3K27ac/me3 by training on all data from all promoter regions at all time-points (Figure 6.6). Calculating the transition probabilities across time between the discovered chromatin states for

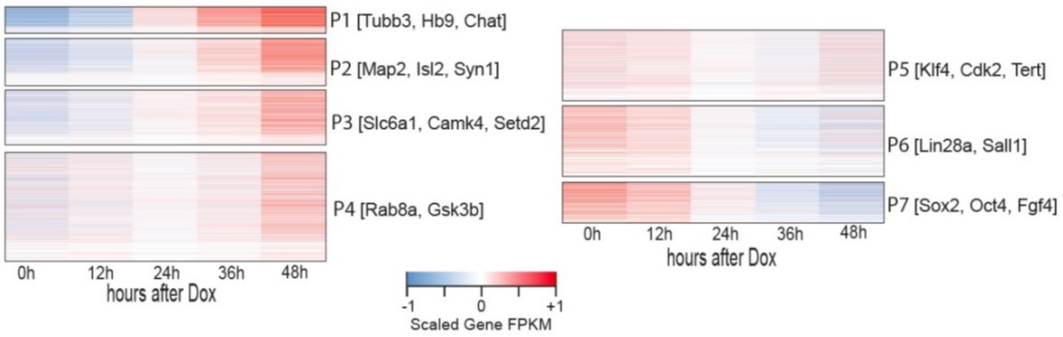


Figure 6.4: Detailed overview of gene expression dynamics for the different up- and down-regulated promoter classes. Gene FPKM values were scaled on the gene level to highlight gene expression dynamics. The height of the heat map of each promoter class is related to the number of genes that are unambiguously assigned to it (genes per class: P1=773, P2=1241, P3=1472, P4=2878, P5=1927, P6=1875, P7=1020, P8=1682, P9=1758, P10=2022, P11=2325). Figure adapted from [155].

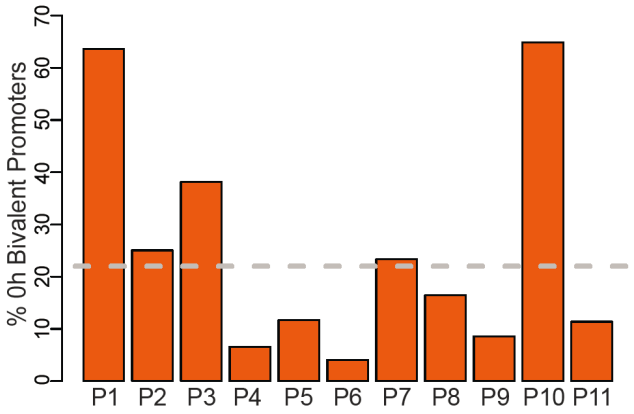


Figure 6.5: Percentage of promoter regions classified as bivalent (H3K4me3-positive and H3K27me3 >H3K27ac) in each promoter group. Dashed gray line indicates the percentage of bivalent promoters out of all promoter regions included in the analysis. Figure adapted from [155].

promoters that started in a bivalent state and transitioned to an active state shows a clear preference for the bivalent chromatin state to transition to a trivalent state followed by a transition to an active chromatin state (Figure 6.6).

This result can be interpreted as transitioning through asymmetric acetylation / methylation on the two H3 histone copies, or the presence of H3K27ac and H3K27me3 on adjacent nucleosomes or as a mixture of cells exhibiting both bivalent and active chromatin. Although none of those possibilities can be confirmed without more sophisticated ChIP experiments such as ChIP-reChIP or Co-ChIP [34], this result points to interesting dynamics of interactions between enzymes setting and deleting H3K27me3 and H3K27ac during bivalency resolution that warrant further investigation.

6.4 Methods[3]

All ChIP-Seq and RNA-Seq data were produced as described in [155] (Sivlia Velasco and Mohamed Ahmed Al-Sayegh / Mazzoni Lab).

Bulk RNA-Seq and ChIP-Seq preprocessing

Expression was quantified from RNA-seq using the Gencode [359] mm10 transcriptome (vM3) and RSEM (parameters: –output-genome-bam forward-prob=0 calc-ci) [63]. RSEM was set to use bowtie1 for read alignments [58]. The geometric average of RSEMs expected FPKM across the biological replicates was used for all further analysis.

All histone modification ChIP-Seq fastq files were aligned to mm10 genome build using bowtie2 [59] with default parameters. After filtering for uniquely-aligned reads that had 2 or less mismatches, potential PCR duplicates were removed using samtools rmdup (parameters: -s) [332]. Resulting BAM files were converted to BED format using bedtools bamtobed command when necessary [333]. For H3K4me1, replicates files were concatenated for all further analysis. JAMM was used to obtain the average fragment length for each experiment [73].

Promoter time-course chromatin state clustering

Promoter regions were defined as -200bp to +2000bp at all annotated Gencode mm10 TSSs (vM3). All overlapping promoter regions were merged regardless of strand to obtain unique non-overlapping promoter regions. JAMM's SignalGenerator script was used to generate depth-normalized, background-subtracted bedGraph files at promoter regions for H3K4me3, H3K27ac and H3K27me3 at 1bp resolution (parameters: -n depth -b 1). The average signal for each histone modification at each promoter region was obtained from those bedGraph files using bedOps bedmap command (parameters: mean).

[3]Text in this section is largely copied directly from [155]

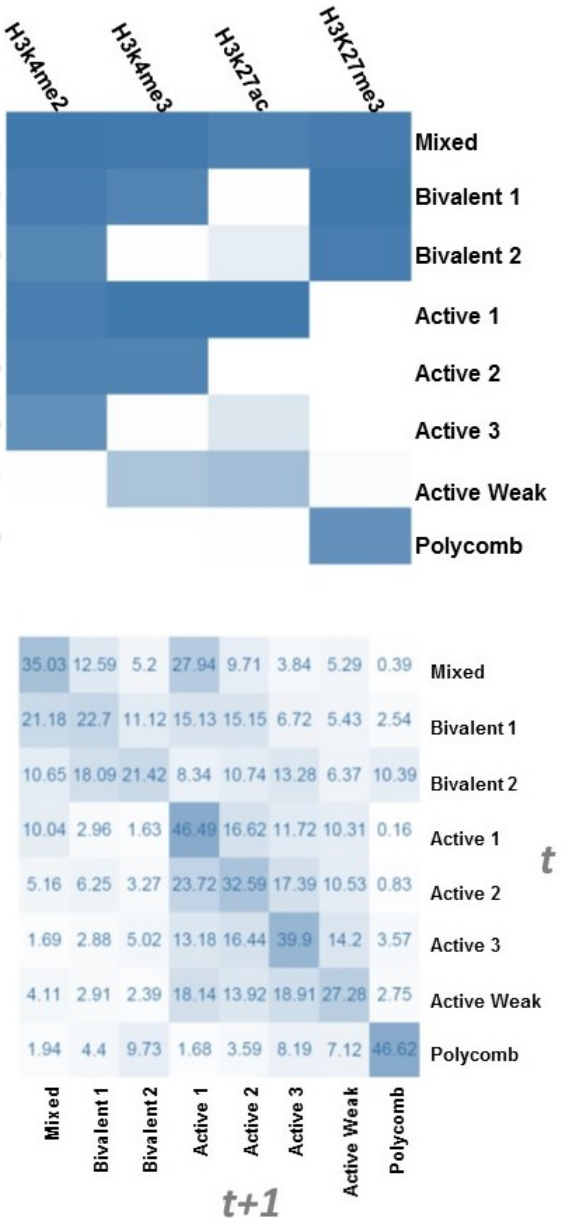


Figure 6.6: Top: Static Chromatin States (see Chapters 4 and 5) learned on annotated promoter regions genome-wide using H3K4me2, H3K4me3, H3K27ac and H3K27me3. Bottom: Transition frequencies (expressed as percentages) between the different chromatin states across time in promoters that start in a bivalent state and transition to an active state (see Methods). Bivalent chromatin state is most likely to transition into a mixed trivalent H3K27ac/H3K27me3/H3K4me3 state which is in turn most likely to transition into an active H3K4me3 / H3K27ac state.

Each histone modification is quantile normalized across all time points using normalizeQuantile command from the limma R package with default parameters [361] (see Chapter 2 for an explanation of quantile normalization). All promoter regions that have lower than background levels for all clustered histone modifications at all time-points are removed from further analysis. Background for each histone modification is defined as the arithmetic mean of its signal across all time-points and all promoter regions. This yields 22302 promoter regions. The log2 fold-change between each two consecutive time-points for each histone modification at each promoter region is calculated after adding a pseudocount of 1 to all values.

To obtain combinatorial time-course clusters of promoter regions based on multiple histone modification datasets across multiple time points, we designed a Bayesian Network (BN) [20] model with a conditional Gaussian probability distribution [357] (see above). For NIL differentiation, we opted to model the chromatin trajectory as a simple linear chain without any branches as predicted via our single-cell RNA-seq analysis (not shown) [155]. To learn point estimates of the parameters of the Bayesian Network model, we use the Expectation-Maximization algorithm for Bayesian Networks implemented in the MATLAB Bayesian Network Toolbox (BNT) [362]. The EM algorithm is initialized via MATLABs kmeans command (parameters: distance, cityBlock Replicates, 15 MaxIter 300). The junction-tree inference engine implemented in the BNT toolbox is used to assign each promoter region a probability of belonging to each of the learned clusters of chromatin trajectories. Each promoter region is assigned to the cluster with the highest probability. To determine the final clustering of the data, we trained our model on all promoter regions available for clustering, then assigned each promoter region to the cluster with the highest probability. (Figure 6.3) shows the quantile-normalized ChIP-Seq values (see above), after linearly scaling the values to ensure histone modifications are comparable, averaged over all promoter regions that belong to a given cluster. Scripts used for preprocessing, clustering and plotting are available at: https://github.com/mahmoudibrahim/timeless

The corresponding RNA-Seq FPKM plots are made using the default R boxplot function on the logarithm of RSEM FPKM values after adding a pseudo-count of 1. Outliers are not displayed. Genes that have multiple promoter regions (due to alternative promoters or alternative transcripts) assigned to different promoter chromatin clusters were excluded from the RNA-Seq plots and from Gene Ontology analysis. The corresponding gene expression heat maps (Figure 6.4) are made on the same FPKM values after centering the expression of each gene at zero, by subtracting the mean across time for each gene from each time-point of that gene.

Chromatin State Analysis

Chromatin states were learned using the same piplines described in Chapter 5 on all promoter regions using histone modification data from all time-points. To produce the transition frequency heatmap, the number of time each state was followed by another across time was counted at 10 basepair resolution. Promoter regions for counting transition frequencies were based on an old version of the time-course promoter clustering

presented in this chapter (not shown).

6.5 Discussion

Generalized methods for summarizing multiple genome-wide data sets across multiple time points require models that can encode time dependencies accurately but without further significant assumptions about how the data should behave. A previous model to co-cluster multiple histone modification data across time [125] assumed a linear differentiation trajectory with only two possible states for each genome region type (see Chapter 2). Obviously the assumption of a linear trajectory does not hold when one is analyzing multi-lineage differentiation trajectories. The second assumption of only two states per genomic region, or in fact the assumption of any certain number of states that is appropriate for all regions, also seldom holds. In this analysis of promoter regions chromatin dynamics during NIL differentiation, we show that promoters go through multiple different states during differentiation.

The Bayesian Network we propose to cluster hisone modifications over time is applicable to arbitrarily complex lineage differentiation trees and does not assume any certain number of states the promoter regions would have to transition through. Instead, the model focuses on the histone modifications signal time dynamics in terms of modeling time dependencies between the fold-change values, assuming that the change between any two time points is related to the previous two time points via simple linear regression. In a sense, one can think of this as trying to model the "acceleration" of the signal across time. The model outputs directly interpretable clusters of promoters where each cluster is explained by the combined time trajectories of multiple histone modifications. "[The model is also] seamlessly extended to as many histone modification data sets as necessary, potentially even if the time-points assayed do not match, as well as any data type that can be represented as log fold-change values." [155]

In static snapshots of cell chromatin environment, histone modifications correlate with expression and with each other (see Chapters 3 and 5, [218, 219, 318]). In this chapter, we showed clearly that histone modification time trajectories also explain gene expression time trajectories to a great extent. However, in the Bayesian Network model employed here, histone modifications are modeled as independent of each other given the cluster assignment. It is worth considering what the independence assumption in this model means. Since we model fold-change values and the dependencies between them across time, we are actually assuming that the *change* in the value of a histone modification across time is independent of the change of another across time. This is different from assuming that histone modifications' counts are independent of each other in static snapshot data sets (the assumption made by models like ChromHMM [315, 316] and Segway [317], see Chapter 4). Nevertheless, the assumption of conditional independence between the rates of change of different histone modifications may still not hold if there are direct or indirect dependencies between the enzymes and the pathways that lead to deposition and deletions of the histone modifications modeled ([363], also see Chapter 3). We had tried regressing multivariate Gaussian

nodes in earlier versions of the model, thereby considering the covariance between the histone modification fold-change values explicitly, but this led to less stable and less meaningful results (not shown). This was potentially due to the large increase in the number of parameters required and/or due to the potential numerical instability of the junction-tree inference algorithm implemented in the Matlab BNT toolbox for conditional Gaussian Bayesian networks. More numerically stable inference algorithms for conditional Gaussian Bayesian networks are described in [364] and [365]. It would be useful to attempt a model where histone modifications dependencies are explicitly considered, potentially still keeping the model as sparse as possible for example by parameter tying.

In Chapter 4, we attempted an explicit representation of ChIP-Seq replicate information and show that this indeed often leads to more accurate peak finding [73]. The Bayesian network described here does not explicitly take advantage of replicates. A possible idea to include replicate information is use replicates to influence the estimates of the variance parameters of the Gaussian nodes thereby determining promoters whose trajectories show "significant" change versus others that might then get assigned to an "ambiguous" cluster, similar to the approach followed in [111]. These ideas can build directly on previous work on pooling data for variance estimation [90–92] and time-course gene expression clustering (see Chapter 2)

Understanding promoter chromatin dynamics during differentiation is essential to understanding how dynamic gene regulatory pathways operate as cells change their identity. We took advantage of a highly efficient and homogeneous system of directed differentiation where pluripotent cells are programmed to spinal motor neurons within 48 hours to study promoter chromatin dynamics as genes get switched on and off. Using time-course ChIP-Seq and the Bayesian Network model to co-cluster multiple histone modifications over time, we show that (1) although bivalent promoters appear to respond faster to activation, bivalent chromatin state is not required for subsequent activation, (2) there are multiple distinct dynamics of activation and repression that we could not entirely explain by the initial chromatin state and that (3) bivalent promoters are likely resolved to active promoter chromatin state by transitioning through a trivalent H3K27me3 / H3K27ac / H3K4me3 chromatin state.

It is interesting to think about the establishment and resolution of bivalent chromatin and how this relates to the establishment and resolution of bona fide repressed states. If during development bivalent domains are established at some developmental promoters but not others and if this really does affect promoter response time to activation signal, it might imply a regulatory network that can pre-determine the future required times and levels of promoter activation regardless of immediate downstream events. *in vitro* differentiation systems starting with ESCs might not faithfully represent *in vivo* systems in terms of pre-establishment of bivalent chromatin [229]. Studies on the chromatin landscape in early embryonic development *in vivo*, such as [261] and [259], might help resolve those questions.

From an engineering point of view, we are interested in understanding how promoter chromatin behaves when cell fate conversion is forced within a short span of

time. Observing a trivalent chromatin state through which bivalent promoters transition to an active chromatin state might indicate a regulatory bottleneck through which promoters transition to reach the active state. Interactions between the PRC complex, histone acetyltransferases and the polymerase preinitiation and elongation complexes (see Chapter 3) should be probed in order to understand bottlenecks to bivalent chromatin resolution in kinetic terms.

Chapter 7

Enhancer Dynamics during Motor Neuron Programming

7.1 Contribution Statement

Results in this chapter are based on work in the following publication: [155]. Similar results to those included in this chapter are provided in [155] using a different approach in collaboration with Akshay Kakumanu (Mahony Lab / Pennsylvania State University).

Contributions to Chapter Results: Silvia Velasco (Mazzoni Lab / New York University) produced all the ChIP-Seq data. Akshay Kakumanu (Mahony Lab / Pennsylvania State University) performed the transcription factor ChIP-Seq peak calling upon which the binding site chromatin clustering is based. Akshay also produced the ChIP-Seq peak calling used for Ngn2 analysis in Appendix II. Antje Hirsekorn (Ohler Lab / Max-Delbrueck-Center, Berlin) produced the ATAC-Seq data provided in Appendix II. Mahmoud M Ibrahim performed all other analysis and results provided in this chapter and in Appendix II.

7.2 Introduction

In this Chapter, we will further examine the NIL-directed spinal motor neuron differentiation system ([155, 160], see Chapters 3 and 6) but from the point of view of distal regulatory elements. Using ChIP-Seq of the NIL factors during the time-course of the differentiation process, we can determine the binding sites of those factors and then examine the changes in their chromatin environment using histone modification time-course ChIP-Seq. Since the vast majority of NIL binding sites occur at distal regions [155], those sites can be assumed to represent enhancers or at least distal regulatory elements in general and I will frequently use the term "enhancer" to refer to such regions. As such, the viewpoint employed here is that clustering time-course histone modification dynamics would directly allow for identification of different groups of enhancers that we hypothesize would regulate different groups of promoters.

In direct programming of cell fates, induced transcription factors are frequently selected to be terminal factors that can directly activate the desired cell fate bypassing regular developmental progenitor stages (see Chapters 3 and 6). It is not clear however how such programs operate on the gene regulatory level. It was previously suggested that the induced factors would recognize a starting chromain state in closed chromatin and then stably bind to those sites [366]. However, those results were obtained using population ChIP-Seq in an inefficient trans-differentiation system confounded by the induction of unintended regulatory networks [167] (see Chapter 3). Hence, this result is not necessarily accurate. In an entirely different system which trans-differentiates fibroblasts to pluripotent cells, a dynamic transcription factor network was observed where the induced factors bind cooperatively to a broad set of enhancers and then iteratively refine the binding sites [166]. However, this was also observed using cell population-based ChIP-Seq in an inefficient trans-differentiation system.

The high efficiency of the NIL directed differentiation system offers an opportunity to study transcription factor binding dynamics that lead to efficient direct programming and perhaps resolve this apparent contradiction. Since Isl1 and Lhx3 are often cobound [155, 160], while Ngn2 cobinds with Isl1 and Lhx3 in only a minority of sites ([155], not shown), we will consider Isl1 and Lhx3 binding below and Ngn2 binding separately in Appendix II.

7.3 Results

7.3.1 Enhancers Time-course Chromatin States

We started by applying the Bayesian Network clustering model developed in Chapter 6 to cluster the time-course dynamics of H3K4me2, H3K4me1 and H3K27ac data sets at all Isl1/Lhx3 time-course binding sites and obtained five clusters (Figure 7.1, data produced by Silvia Velasco / Mazzoni Lab. Isl1/Lhx3 transcription factor binding sites were defined by Akshay Kakumanu / Mahony Lab, see Methods). E1 cluster starts in an active state at the embryonic body (EB) stage at 0h, represented by high levels of active histone modifications and rapidly loses its activation signal. In contrast, E2, E3 and E4 clusters all start in an low activity state and gain activation signals but with different time-course behaviors. E2 regions become active only briefly at 12h while E4 becomes active only later during differentiation. E3 enhancers become active early at 12h and remain active gaining higher levels of activation until the end of the differentiation. Finally, E5 enhancers start at an inactive state and remains so during differentiation.

7.3.2 Enhancer Dynamics Correlate with Promoter Dynamics

We have now established that distal regulatory regions bound by Isl1 and Lhx3 show different enhancer activation (and repression) dynamics. We have also established in Chapter 6 that different promoter regions have different activation and repression

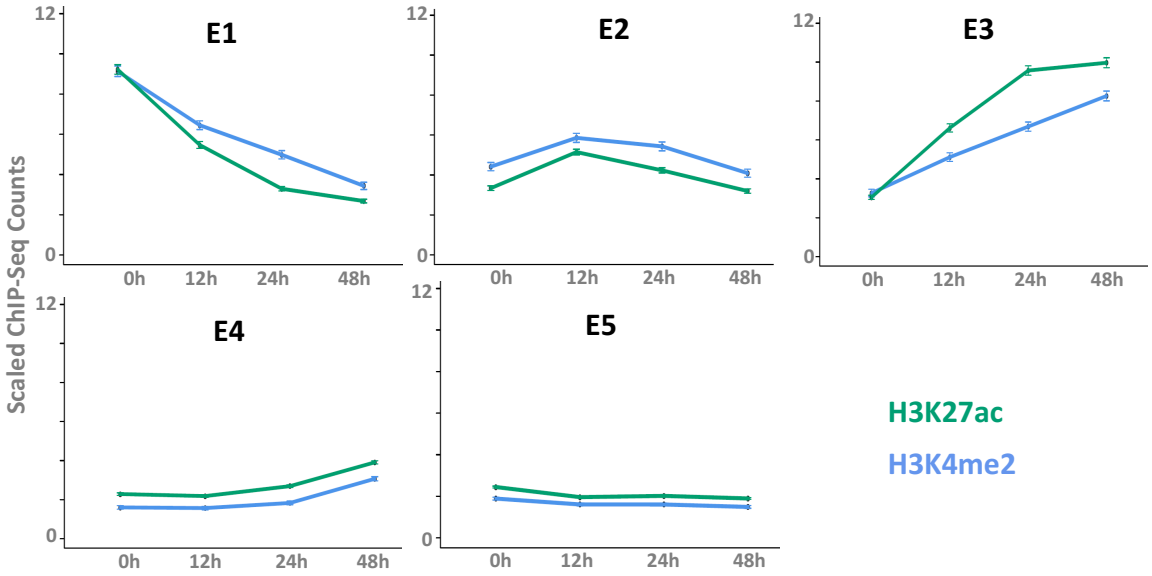


Figure 7.1: 5 Clusters obtained from distal Isl1/Lhx3 binding sites based on their combinatorial histone modification time-course behavior.

dynamics. How do enhancer dynamics relate to promoter dynamics?

To answer this question, one has to assign distal binding sites to promoters with a method that is agnostic in terms of matching binding site activity to promoter activity. In the absence of genome conformation data, the most straightforward way this might be possible is to rely on genomic distance to assign enhancers to promoters. An obvious method is to assign each binding site to its closest promoter, establishing a one-to-many relationship of promoters to binding sites. We found this model (especially when restricted to a 100kb distance) to enrich for positive correlations between binding sites and promoters ([155], not shown). This assignment generates a matrix of transcription factor binding site groups / promoter region groups association frequencies. To estimate whether a certain enhancer group is matched to a certain promoter group more often than what would be expected by random, one needs a suitable null model. To obtain a suitable null model, the association frequency matrix was randomized 100,000 times but requiring that row sums and column sums remain the same. Plotting the log2 fold-change values of the observed association frequency matrix to the averaged randomized matrix reveals the enrichment/depletion of association compared to the randomized matrix (Figure 7.2).

This log2 enrichment/depletion matrix reveals four main key observations (Figure 7.2): (1) E1 binding sites are depleted from up-regulated promoters but enriched in down-regulated promoters. This raises two possibilities: the NIL factors bind 0h active enhancers and that this binding might be related to down-regulation of stem

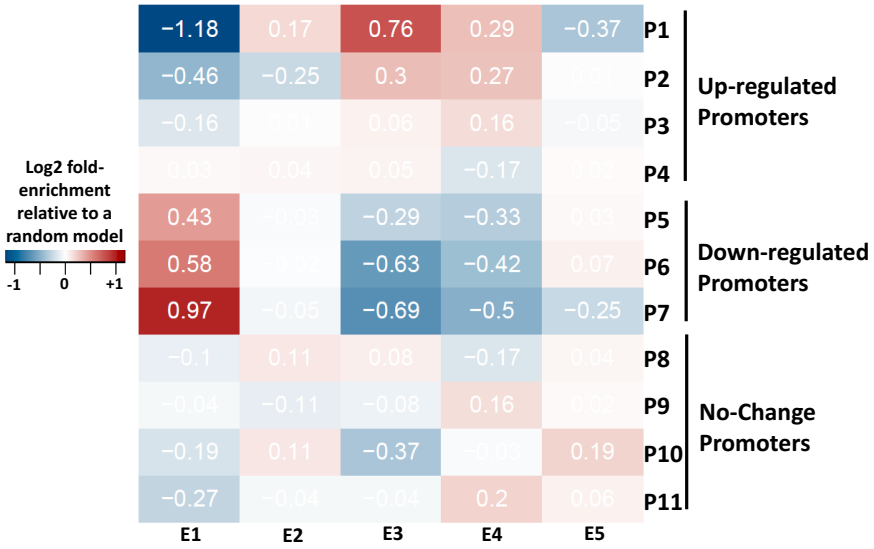


Figure 7.2: Matrix shows log2 fold change values of binding site / promoter association frequency relative to a random model of association. Positive values indicate enrichment of a binding site group with a promoter group, negative values indicate depletion.

cell and cell cycle genes or that this binding is off-target binding, (2) E2 sites show no enrichment for any promoter cluster except for a weak enrichment in up-regulated promoter group P1, possibly indicating abortive or unsuccessful binding of the NIL factors, (3) E3 and E4 sites are enriched in up-regulated promoters with faster more strongly activated promoter group P1 showing the strongest enrichment followed by P2, followed by P3, which indicates a correlation between enhancer activation dynamics and promoter activation dynamics and (4) E5 binding sites are depleted from strong up-regulated P1 promoters and strongly down-regulated P7 promoters and do not show any enrichment in other promoter classes except P10 promoter group which is the group that fails to resolve bivalency to activation or full repression, indicating that those binding sites might be abortive unsuccessful off-target sites.

This raises a picture of a highly targeted, albeit multi-step, process of directed differentiation. But one central question remains: Why do Isl1 / Lhx3 binding sites show different enhancer activation / decommissioning dynamics?

7.3.3 Transcription Factor Cooperativity Explain Enhancer Dynamics

To answer this question, we started by examining the transcription factor ChIP-Seq levels for Isl1/Lhx3 factors at the different classes (Figure 7.3). Two conclusions are obvious: (1) as expected Isl1 and Lhx3 dynamics correlate to a great extent (see Chapter 3 and [155]) and (2) the dynamics of Isl1 and Lhx3 mirror closely those of

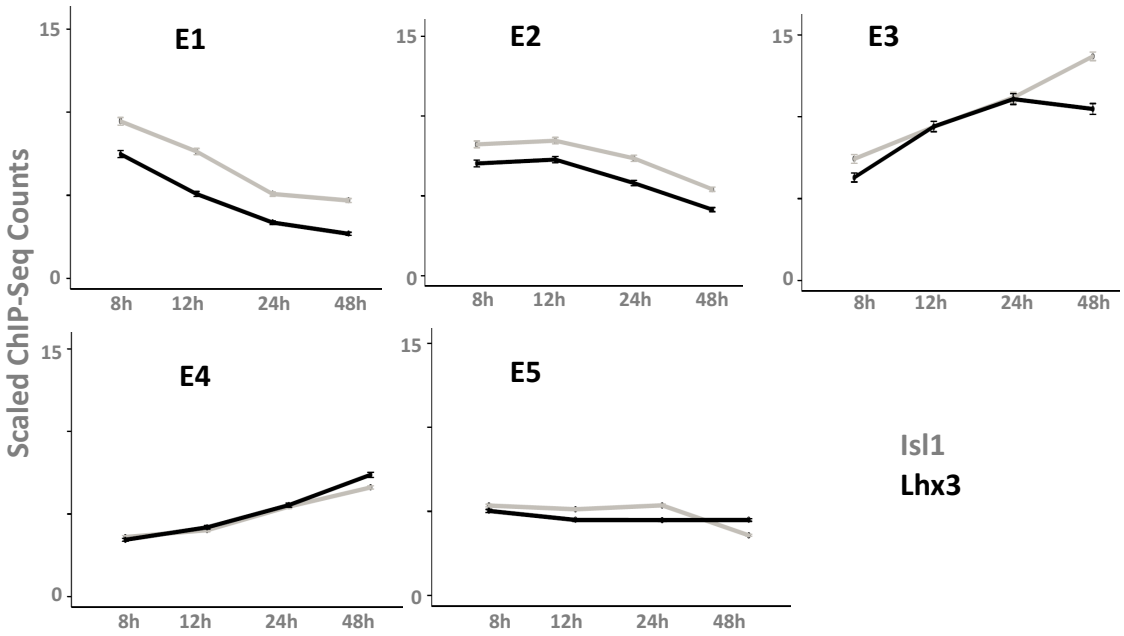


Figure 7.3: Isl1 and Lhx3 time-course histone modification signal for the 5 Clusters obtained from distal Isl1/Lhx3 binding sites histone modification time-course behavior.

histone modifications measuring enhancer activity. Therefore, the time trajectories of enhancer activation dynamics measured by histone modifications are most likely the result of changes in the binding dynamics and binding sites of Isl1 and Lhx3.

Why do Isl1 and Lhx3 change their binding sites during differentiation? Changes in Isl1 and Lhx3 binding might be explained by a passive model that relies on the initial chromatin landscape present when Isl1 and Lhx3 are first induced (at 0h): Isl1/Lhx3 initially bind 0h active and accessible sites (probably pluripotency enhancers) opportunistically in an off-target fashion. And as Isl1/Lhx3 start accessing sites that were initially closed but have more Lhx3-favorable binding site sequences, the binding shifts from the initial off-target sites to the more favorable on-target sites. This model can explain E1 enhancers and E3 enhancers but fails to explain E2 and E4 enhancers which are activated only briefly early and later during differentiation in a manner that seems to be intimately linked to promoter activation and repression dynamics. Although we only induce three factors, they are of course not the only factors active during the differentiation process. At the very least the pluripotency network transcription factors are active at the early phases of the differentiation process, and in addition, Isl1 and Lhx3 might activate other factors downstream. Furthermore, Ngn2, a pioneer transcription factor (see Chapter 3), is expected to activate other transcription factors downstream of it. For example, Ngn2 has been shown to cooperate with Isl1 and Lhx3 to activate Hb9 (an important spinal motor neuron gene) enhancers [367, 368]. We set out to investigate the presence of other potential factors that might cooperate with the Isl1 and

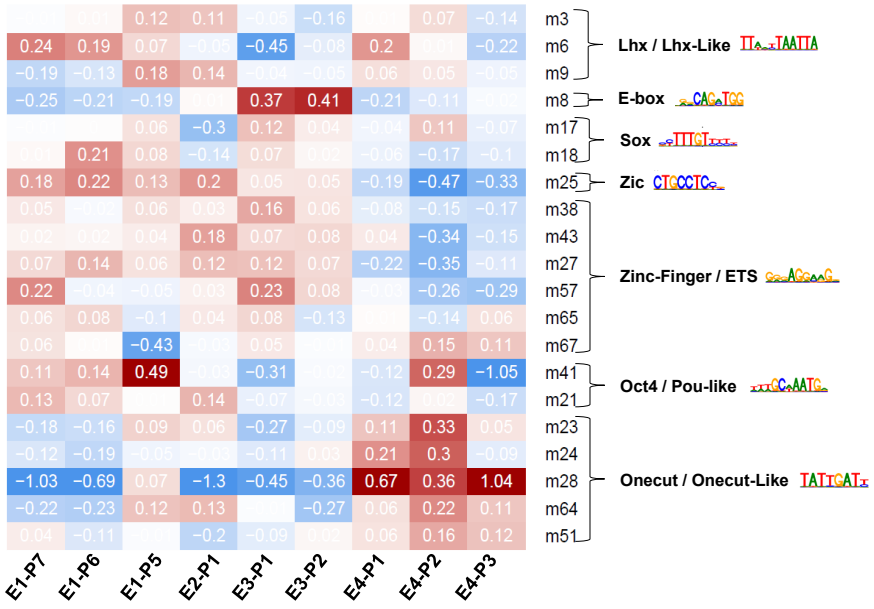


Figure 7.4: Matrix shows log2 fold change values of fraction of sites containing a certain motif relative to the fraction of all sites containing that motif. Enhancer-Promoter group indicates Isl1/Lhx3 sites that are assigned to a certain enhancer group based on enhancer chromatin dynamics (example: E1) and are assigned to a certain promoter group (example: P1). Motif logos indicated are chosen from the respective motif group indicated. A list of all motifs and their logos is available in Figure 7.6.

Lhx3.

To understand why the Isl1/Lhx3 switch their binding sites during differentiation, we focused on the binding site groups that are enriched in up-regulated and down-regulated promoters and further split each binding site group by which promoter group the binding sites are assigned to. Thus we obtained nine enhancer-promoter pair groups that are classified based on their own enhancer time-course histone modification trajectories and the histone modification trajectories of the promoters they potentially regulate. We searched for motifs in all those sites *de novo* using XXmotif [369], then we quantified the percentage of motif occurrence in each binding site / promoter pair group using FIMO motif scanner [327] (see Methods). To quantify the enrichment or depletion of a certain motif in each enhancer-promoter pair group relative to the overall abundance of that motif in all binding sites, we plot the log2 fold change of the fraction of sites containing a motif in an enhancer-promoter group to the corresponding fraction in all sites put together (Figure 7.4).

Overall, in addition to the expected NIL motifs (Lhx Lim homeodomain and bHLH factors E-box motifs), we observe additional motifs belonging to other homeodomain factors like Oct4 as well as Sox factors, factors from the Zic family, factors from the Onecut family and Zinc-finger or ETS-like motifs (Figure 7.4). This provides evidence for high transcription factor cooperativity potential encoded in the genome at Isl1/Lhx3

binding sites. Further, not all motifs are equally enriched in all enhancer-promoter groups. E1 enhancer groups are enriched in Oct4, Zic, Sox and Lhx motifs but depleted from Onecut and E-box motifs. In contrast, E4 enhancer groups are enriched in Onecut motifs but generally depleted from Oct4, Zinc-finger, Zic, Sox and E-box motifs, while E3 enhancer groups are enriched in E-box and Zinc-finger motifs, depleted from Oct4 and Onecut motifs and generally not enriched nor depleted in Lhx motifs. Finally, E2 group is not enriched in Sox or E-box motifs but enriched in Zic and Zinc-finger motifs. Therefore, the different enhancer groups E1, E2, E3 and E4, although all bound by the Isl1/Lhx3 factors, are distinguished by different combinations of motif groups indicating potential cooperativity between the Isl1/Lhx3 factors and other factors that are either expressed at 0h as part of the pluripotency program (Oct4, Zic, Sox2) or potentially expressed later during differentiation (Onecut, other Sox factors).

When put together with Isl1/Lhx3 binding dynamics, a picture emerges where a transcription cooperativity model, in contrast to the passive model explained above, becomes responsible for Isl1/Lhx3 binding dynamics. In this model, early binding at 8h-12h occurs in E1, E2 and E3 sites. E1 sites binding to 0h active sites associated with pluripotency factors like Oct4, Sox(2) and Zic motifs is associated with enhancer decommissioning and the binding in those sites is quickly lost. E3 sites are initially inactive at 0h but become bound likely in cooperation with Ngn2 [367, 368], and maintained potentially with the help of Sox factors and Zinc-finger factors expressed early on when differentiation starts. E2 sites are not productively activated and maintained potentially due to the less frequent presence of E-box and/or Sox motifs. The absence of E-box, Sox and Zinc-finger motifs in E4 sites can also explain E4 sites and their dynamics. One can hypothesize that those sites require Onecut expression (Onecut is expressed later during differentiation [155], not shown) to activate those sites and enable productive binding of Lhx3. Therefore, an active model where factors expressed early during differentiation, including pluripotency factors, as well as other factors expressed later during differentiation are responsible for the transcription factor binding dynamics of the NIL factors which were directly induced. This model is further supported by the ChIP-Seq enrichment of Oct4 at 12h, Ngn2 at 12h and Onecut2 at 48h at the different enhancer groups (Figure 7.5).

Finally, one can notice that within one enhancer group, there are different preferences for different motif combinations. For example, E1 sites regulating P7 and P6 are distinguished from each other by preferences for motifs m3, m6, m8, m18 and m57 while E3-P1 sites are distinguished from E3-P2 sites by stronger Sox and Zinc-finger enrichment (Figure 7.4). This result might be difficult to interpret though without further in depth analysis of motif dependencies and degeneracy because it is hard to infer which exact factor is differentially enriched.

7.4 Methods

All ChIP-Seq data was produced as described in [155] (Silvia Velasco / Mazzoni Lab).

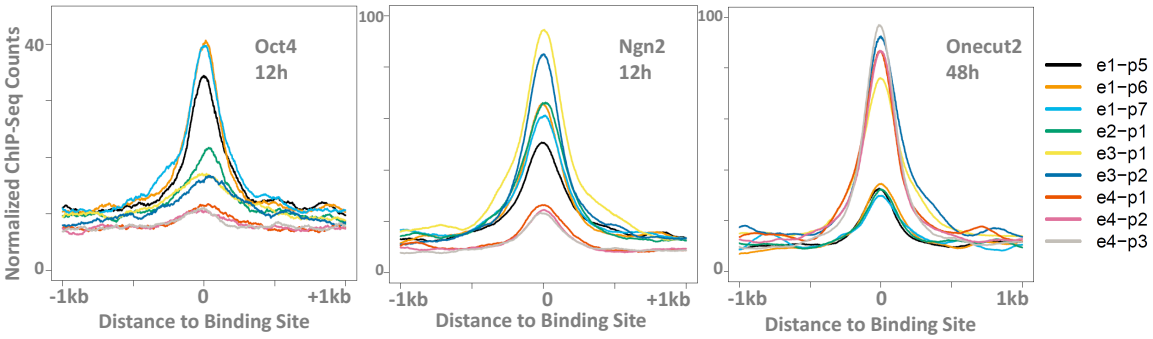


Figure 7.5: Plots show average normalized ChIP-Seq signal for Oct4 at 12h, Ngn2 at 12h and Onecut at 48h at the Isl1/Lhx3 sites belonging to the different enhancer-promoter groups.

Binding Site Chromatin Clustering

Single-basepair Isl1 and Lhx3 binding sites at 12h, 24h and 48h were obtained using multiGPS [370] as described in [155] (Akshay Kakumanu / Mahony Lab).

Binding sites were then extended by the average fragment length calculated by JAMM [73] + 150bp in each direction (total 330 bp extention in each direction). All overlapping sites are merged to obtain unique non-overlapping regions. Signal for all histone modifications and transcription factor ChIP-Seq and clustering using the conditional Gaussian Bayesian Network was done as described in Chapter 6.

Promoter-Enhancer Assignment

"Each transcription factor binding site was assigned to its closest promoter region, requiring that the distance is 100kb or less. This generates a matrix of transcription factor binding site / promoter regions association frequencies, expressing a one-to-many relationship of promoter regions-to-transcription factor binding sites. To obtain a suitable null model , the association frequency matrix was randomized 100,000 times but requiring that row sums and column sums remain the same, using the function permatfull in the R package vegan (parameters: fixedmar=both burnin=1000 time=100000). Values plotted in Figure 2 are the log2 fold-change values of the observed association frequency matrix to the averaged randomized matrix. Values higher/lower than zero indicate enrichment/depletion of association compared to the randomized matrix." [155]

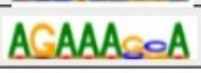
Lhx / Lhx-Like	M3	
	M6	
	M9	
E-box	M8	
Sox	M17	
	M18	
Zic	M25	
Zinc-Finger / ETS	M38	
	M43	
	M27	
	M57	
	M65	
	M67	
Oct4 / Pou-Like	M41	
	M21	
Onecut / Onecut-Like	M23	
	M24	
	M28	
	M64	
	M51	

Figure 7.6: List of motifs enriched in Isl1/Lhx3 binding sites. See Figure 7.4.

De novo Motif Search

Enhancer groups that show specific enrichment with a promoter group were selected for further analysis. All sites were concatenated and truncated to 201 basepairs around the midpoint (the binding site). De novo motif finding was then started using XXmotif [369]. The list of motifs obtained was filtered to keep only motifs that are present in at least approximately 5 % of all sites as quantified by XXmotif. FIMO [327] was then used to obtain the frequency of each motif in the each group of enhancer cluster / promoter cluster pair. Motif identities were determined with the help of TomTom web server on the "Vertebrates (in vivo and in silico)" database [371].

ChIP-Seq Average Plots

"To generate average plots for histone modification data, ChIP-Seq replicate experiments were concatenated and converted to bigwig files using deepTools bamCoverage at 10bp resolution (parameters: –normalizeUsingRPKM [335], using the average fragment length predicted by JAMM [73]. bedGraph files generated by JAMM were used for ATAC-seq. deepTools computeMatrix [335] was then used to generate the counts at the regions of interest. ChIP-Seq input was subtracted from the ChIP-Seq data at each binding site and each position and values lower than zero were considered zero. The arithmetic mean at each position is then plotted in R. All heat maps were plotted using the heatmap.2 function in the gplots R package." [155]

7.5 Discussion

Transcription factors participate in gene regulation by recognizing DNA sequence motifs at distal regulatory regions and at promoters. Combinations of motifs at regulatory sites underlie the regulatory computational space available for a cell to tune its gene expression. In this chapter, we explored the enhancer landscape during directed differentiation of spinal motor neurons. Isl1 and Lhx3 change their binding sites during differentiation leading to different groups of enhancer activation and decommissioning dynamics, which correlate with chromatin dynamics at the promoters they regulate. Therefore, the forced expression of Ngn2, Isl1 and Lhx3 leads to a highly coordinated multi-step time-varying regulatory network that takes advantage of other secondary transcription factors to fine tune Isl1 and Lhx3 binding to distal regulatory regions. The use of such secondary factors is possible due to co-occurrence of the Lim-homeodomain Lhx motif with other secondary motifs like Oct4, E-box and Onecut, thus enabling cooperativity between those factors and Isl1/Lhx3. Because those secondary factors are expressed at different times during the differentiation process, Lhx3 can cooperate with those factors only as they become expressed, leading to dynamic Lhx3 binding time trajectories during the short span of 48 hours of motor neuron programing. This result is in contrast to previous results obtained from an analysis of a fibroblast-to-neuron direct programming system where induced factors target closed chromatin with a specific chromatin state and remain bound [366]. This

system is inefficient and therefore conclusions about transcription factor binding from population ChIP-Seq data are confounded by cells that did not properly differentiate [167]. Since the NIL system is efficient and homogeneous [155, 160], it is possible that a dynamic transcription factor binding behavior is the more common theme in trans-differentiation direct programming systems.

The dynamic enhancer landscape observed during motor neuron programming highlights the complex regulatory logic of enhancer-promoter regulation: each enhancer responds to multiple cooperating factors and each promoter responds to a combination of enhancers regulating its dynamic behavior. This indicates a complex picture of time-varying enhancer-promoter networks. The contribution of enhancer dynamics to promoter dynamics relative to the contribution of promoters' own chromatin bottlenecks (Chapter 6) is difficult to determine and quantify. In this chapter, one observes a strong correlation between enhancer dynamics and promoter dynamics but we also observed a tendency for promoters with the highest rate of activation to have a bivalent state at 0h (Chapter 6). Therefore, a plausible hypothesis could be that both aspects contribute to promoter dynamics and that different promoters might respond differently to different enhancer combinations depending on their initial chromatin state. Moving from gene-gene network models to models where enhancers are represented explicitly can help identify network motifs and common aspects of enhancer-promoter dynamics in differentiation and development, and generate testable hypothesis for future experiments.

Traditionally, researchers were focused on gene regulatory network inference from gene expression data alone and when transcription factor binding data became available, this view was maintained and models were only refined to reflect more accurate gene-gene interaction information (see Chapter 2). Therefore, with the availability of studies like this one where gene expression is engineered in a certain clear direction and chromatin regulation is measured over time, it should be possible to infer and build gene regulatory networks where enhancers, their chromatin regulation and their effect on promoter chromatin are directly represented. However, we still fell short here of simulating an integrated enhancer-promoter regulatory network because of two main reasons. First, such a regulatory network model greatly expands the number of regulatory entities to be modeled and representing chromatin regulation directly also greatly expands the number of parameters each entity (ie. enhancer, promoter...etc.) would need to be modeled. Therefore, it would probably require innovative novel modeling frameworks and significant computational resources to fit and simulate such models. Second, it is not entirely clear whether the time resolution that this and other studies employ is sufficient to infer such networks. It is naturally a massive investment in cost to produce similar data sets with a higher time resolution.

Chapter 8

Discussion and Outlook

Genome-wide Time-course Clustering Models

The readout of transcription regulation is often gene expression as measured by RNA-Seq, assaying the steady-state of stable RNA in the cell. Hence, work on inference of gene regulatory networks has often relied on gene expression data ignoring direct representation of the role of enhancers in gene regulation. These approaches also ignore RNA processing and RNA degradation, processes that cause degeneracy between *chromatin-level* transcription regulation and steady-state RNA levels.

Throughout the thesis, we argued for the utility of gene regulatory networks where enhancers are represented as separate entities that interact with each other and with gene promoters. Although this work does not include an integrated method to automatically infer such regulatory networks from chromatin regulatory events, we have made several steps toward that goal. In Chapters 6 and 7, we adopted a representation where all enhancers and promoters are considered as separate independent entities acting in concert during cell differentiation in a dynamic network. To characterize enhancer-promoter associations, we first clustered promoters and enhancers separately via their combinatorial histone modification trajectories using a Conditional Gaussian Bayesian network model that can co-cluster multiple time-course data sets directly, and then measured enhancer-promoter association based on the distance between enhancers and promoters. A more elegant solution would have been to co-cluster enhancer and promoter trajectories together. A possible objective function in that case could be to maximize the information content in the enhancer-promoter association matrix (see Figure 7.2) given the histone modification trajectories at enhancers and promoters. But an algorithm that can do this needs to be thought out carefully to make sure it is guaranteed to converge.

Several other simpler improvements can be applied to the time-course co-clustering model. For example, an input layer can be added to represent transcription factor binding such that transcription factor dynamics are co-learned with histone modification dynamics. This model would for example be able to automatically assign a single transcription factor time trajectory to multiple enhancer and promoter time-course

chromatin trajectories or vice versa, therefore allowing for learning of complex transcription factor enhancer activation and decommissioning logic. Alternatively, an output layer can be added to represent transcription output such as GRO-Seq. This would also allow for degeneracy between chromatin time-course trajectories and gene expression trajectories. For example, the same activation chromatin trajectory can result in an increase in gene expression in some but not all genes due to high rates of RNA degradation in a subset of genes. Note that such a model would potentially contain a discrete child for continuous parents, in which case the discrete child can be modeled as a softmax conditional probability distribution [372]. Alternatively, the input and output layers could be split each with its own discrete class node, requiring a third "overall" class node (see Chapter 6).

Transcription Factor Cooperativity

Transcription factor cooperativity is well-established, for example in motor neuron differentiation Ngn2 interacts with Isl1 and Lhx3 to activate the motor neuron factor Hb9 [367, 368]. We show that this cooperativity is also key to directed programming of motor neurons by forced expression of all three factors in pluripotent cells (Chapter 7). The chromatin environment and the available combinations of transcription factor sequence motifs in distal and proximal regulatory regions confer a dynamic transcription factor binding network on Isl1 and Lhx3, even within the span of only 48 hours ([155], Chapter 7). In addition to the interactions between Isl1/Lhx3 and Ngn2 in a minority of sites, we also identify cooperative binding with Oct4 (a pluripotency factor [373–376]) and Zic factors (linked to pluripotency and neuron differentiation [377, 378]) in E1 Isl1/Lhx3 enhancer group which then loses binding and activity quickly and is associated strongly with down-regulated genes. This is interesting for two reasons: 1) it might indicate a role for Lhx3 in decommissioning stem cell enhancers therefore hastening pluripotency exit and 2) it points to the close arrangement of different homeodomain motifs in stem cell enhancers encoding for transcription factors that are likely not naturally expressed together in the same cell type at the same time. This raises questions on the evolution constraints that led to such an arrangement. Furthermore, we also identify Onecut motifs strongly enriched in E4 enhancer group which is activated only later during differentiation consistent with the expression profile of Onecut genes during differentiation (not shown). Onecut factors are known to be important for neuron differentiation [379, 380] and to affect the expression of the Isl1 gene [379]. This again points to similarities between the transcription regulation program employed during *in vivo* motor neuron development and the direct programming of motor neurons even though direct programming skips the motor neuron progenitor states [160]. A similar in-depth analysis of transcription factor binding and chromatin regulation in a stepwise development-like system is still required, in order to confirm the similarities and delineate the differences.

Promoter Chromatin Dynamics

Transcription factor cooperativity and enhancer chromatin dynamics eventually converge on promoters to either enhance transcription initiation and elongation or to inhibit transcription. Promoters have a complex chromatin environment and can exist in multiple different chromatin state combinations at different time-points during differentiation. For example, bivalent promoters are resolved by an increase in H3K27 acetylation paralleled by a decrease in H3K27 methylation (Chapter 6), which raises a possibility of asymmetrical H3K27ac/me3 state at promoters during resolution of promoter bivalency, and indeed one can observe evidence for this in ChIP-Seq data at intermediate time-points during differentiation (Chapter 6). Owing to the short time span of the differentiation, the NIL system is an ideal system to investigate H3K27ac/me3 dynamics since many promoters switch states at the same time in a homogeneous cell population. This enables studying acetylation/methylation symmetry and the relationship between polycomb and trithorax proteins during bivalency resolution using ChIP-reChIP or other similar protocols such as Co-ChIP [34]. But why is it important to understand promoter chromatin dynamics?

One main question that is not clear is how much enhancer dynamics contribute to changes in transcription quantities versus promoter chromatin. This is rather difficult to disentangle because enhancer cooperation logic is not understood and promoter chromatin dynamics are also not understood. In the NIL differentiation system, enhancer dynamics correlate tightly with promoter dynamics but it is not clear whether the target promoter chromatin state affects the magnitude of promoter activation rate or whether it affects the variance in promoter activation rates. Answering such questions requires a higher time resolution data set (the data we used was at 12h and 24h intervals) and requires accurate assays of nascent transcription during differentiation.

Promoters are further complicated by the frequent presence of divergent transcription from promoter open regions (Chapter 5). We showed that divergently transcribed promoters feature a unique chromatin environment in human cells and that histone modification levels on the +1 and -1 nucleosomes correlate with transcription initiation levels for the sense transcript and the antisense divergent transcript respectively (Chapter 5, [298]). However, when studying promoter dynamics during NIL differentiation, we only considered the downstream histone modification levels (promoter regions were defined as -200bp to +2kb at annotated TSSs). It would be interesting to investigate the concordance between chromatin dynamics on the +1 nucleosome and the chromatin dynamics on the -1 nucleosome and how this relates to transcription rates and response to different enhancer combinations.

ChIP-/DNase-Seq Peak Finding

Investigating the chromatin environment at promoters and enhancers to a resolution high enough to separate flanking nucleosomes from ChIP-Seq, DNase-Seq and ATAC-Seq data requires accurate demarcation of the widths and locations of open regions. To this end, JAMM, a peak finder for high-throughput sequencing data, was developed

(Chapter 4, [73]). JAMM offers several advances in this niche field of peak finding including joint analysis of replicate experiments and accurate demarcation of peak widths. JAMM typically produces a large number of peaks when run with its default parameters, with the idea being that users can then have the choice to threshold the peak list in a way that is meaningful to their use case, or use the full sorted list for programs that require a sorted list with false positives at the bottom of the list, or use replicate reproducibility to determine a confident peak set [331]. This can be considered an advantage over many other peak finders which can not relax the peak finding threshold effectively or can not do so without causing peak finding artifacts. For example, MACS [65] peaks become wider when the threshold is relaxed [1]. Although JAMM can currently threshold peaks automatically via a fold-change cutoff that is automatically calculated from data, in future versions of JAMM, an empirical p-value and an empirical false discovery rate will also be calculated since this is what most researchers are now accustomed to from peak finders.

[1] see IDR web page: https://sites.google.com/site/anshulkundaje/projects/idr

Chapter 9

Conclusion

In this work, three main computational advances were introduced: (1) JAMM [73], a peak finder for ChIP-Seq data that can integrate biological replicates, demarcate accurate peak widths and resolve neighboring narrow peaks (Chapter 4), (2) a high-resolution chromatin state discovery pipeline [298] that uses "semi-binarized" signal obtained via JAMM peaks as input to a Hidden Markov Model with multivariate Gaussian emissions (Chapters 4 and 5) and (3) a Bayesian Network model for co-clustering of multiple time-course high throughput data sets (ie. time-course chromatin states, Chapters 6 and 7, [155]).

Using these tools, we studied transcription regulation on the chromatin level, delineating promoter chromatin environment in human cells and how it relates to divergent transcription as well as promoter and enhancer chromatin dynamics during directed motor neuron programming from mouse embryonic stem cells. We found enhancer chromatin dynamics to be in high concordance with promoter chromatin dynamics and that enhancer chromatin dynamics are the result of a complex network of transcription factor cooperativity between the factors induced (Ngn2, Isl1 and Lhx3) and other factors that are expressed at different time-points during the differentiation process.

As our understanding of chromatin regulatory events and how they relate to transcription output improve, inference of transcription regulatory networks that take into account transcription factor cooperativity logic, enhancer cooperativity logic and how this interplays with promoter chromatin and sequence constraints will become feasible. This will potentially allow for more precise engineering of direct programming differentiation systems.

References

[1] Jingyi Jessica Li, Peter J Bickel, and Mark D Biggin. "System wide analyses have underestimated protein abundances and the importance of transcription in mammals." In: *PeerJ* 2 (2014), e270. arXiv: 1212.0587.

[2] R Maurer, T Maniatis, and M Ptashne. "Promoters are in the operators in phage lambda". In: *Nature* 249.454 (1974), pp. 221–223.

[3] G D Stormo. "DNA binding sites: representation and discovery." In: *Bioinformatics (Oxford, England)* 16.1 (2000), pp. 16–23.

[4] Albrecht Kossel. "Ueber einen peptonartigen bestandtheil des zellkerns." In: *Biological Chemistry* 8.6 (1884), pp. 511–515.

[5] R D Kornberg. "Chromatin structure: a repeating unit of histones and DNA." In: *Science (New York, N.Y.)* 184.139 (1974), pp. 868–871.

[6] K Luger, a W Mäder, R K Richmond, et al. "Crystal structure of the nucleosome core particle at 2.8 A resolution." In: *Nature* 389.6648 (1997), pp. 251–60.

[7] Yahli Lorch, Janice W. LaPointe, and Roger D. Kornberg. "Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones". In: *Cell* 49.2 (1987), pp. 203–210.

[8] Scott A. Lacadie, Mahmoud M. Ibrahim, Sucheta A. Gokhale, et al. "Divergent transcription and epigenetic directionality of human promoters". In: *FEBS Journal* (2016).

[9] B D Strahl and C D Allis. "The language of covalent histone modifications." In: *Nature* 403.6765 (2000), pp. 41–45.

[10] C Benoist and P Chambon. "In vivo sequence requirements of the SV40 early promotor region." In: *Nature* 290.5804 (1981), pp. 304–10.

[11] Julian Banerji, Sandro Rusconi, and Walter Schaffner. "Expression of a ??-globin gene is enhanced by remote SV40 DNA sequences". In: *Cell* 27.2 PART 1 (1981), pp. 299–308.

[12] P. Moreau, R. Hen, B. Wasylyk, et al. "The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants". In: *Nucleic Acids Research* 9.22 (1981), pp. 6047–6068.

[13] T Maniatis, S Goodbourn, and J A Fischer. "Regulation of inducible and tissue-specific gene expression." In: *Science (New York, NY)* 236.4806 (1987), pp. 1237–1245.

[14] B G Pogo, V G Allfrey, and A E Mirsky. "RNA synthesis and histone acetylation during the course of gene activation in lymphocytes." In: *Proceedings of the National Academy of Sciences of the United States of America* 55.4 (1966), pp. 805–12.

[15] Mirsky AE Allfrey VG, Faulkner R. “Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis”. In: *Proc Natl Acad Sci U S A* 51.1938 (1964), pp. 786–94.

[16] Paul S. Kayne, Ung Jin Kim, Min Han, et al. “Extremely conserved histone H4 N terminus is dispensable for growth but essential for repressing the silent mating loci in yeast”. In: *Cell* 55.1 (1988), pp. 27–39.

[17] ENCODE Project Consortium, Bradley E Bernstein, Ewan Birney, et al. “An integrated encyclopedia of DNA elements in the human genome.” In: *Nature* 489.7414 (2012), pp. 57–74. arXiv: 1111.6189v1.

[18] A Dempster, N Laird, and D Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *J Roy Stat Soc Ser B* 39.1 (1977), pp. 1–38.

[19] J D Banfield and A E Raftery. “Model-based Gaussian and non-Gaussian clustering”. In: *Biometrics* 49.3 (1993), pp. 803–821.

[20] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Vol. 19. Cambridge U.K. ;;New York: Cambridge University Press, 2000, pp. 675–685.

[21] Regine Kolinsky Jose Morais. *Migrations in Speech Recognition*. Vol. 11. 6. Morgan Kaufmann Publishers, 1996, pp. 611–620.

[22] Andrew J Viterbi. “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm”. In: *IEEE Transactions on Information Theory* 13.2 (1967), pp. 260–269.

[23] Alexander V. Lukashin and Mark Borodovsky. “GeneMark.hmm: New solutions for gene finding”. In: *Nucleic Acids Research* 26.4 (1998), pp. 1107–1115.

[24] Joseph K Pickrell, Daniel J Gaffney, Yoav Gilad, et al. “False positive peaks in ChIP-seq and other sequencing-based functional assays caused by collapsed repeats in the reference genome”. In: *Bioinformatics* 27.15 (2011), pp. 5–7.

[25] Haitham Ashoor, Aurélie Hérault, François Radvanyi, et al. “HMCan a tool to detect chromatin modifications in cancer samples using ChIP-seq data”. In: *Bioinformatics* 29.23 (2013), pp. 2979–2986.

[26] Yuval Benjamini and Terence P. Speed. “Summarizing and correcting the GC content bias in high-throughput sequencing”. In: *Nucleic Acids Research* 40.10 (2012), e72. arXiv: NIHMS150003.

[27] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, et al. “Counting absolute numbers of molecules using unique molecular identifiers”. In: *Nature Methods* 9.1 (2011), pp. 72–74.

[28] Y Chen, N Negre, Q Li, et al. “Systematic evaluation of factors influencing ChIP-seq fidelity”. In: *Nature methods* 9.6 (2012), pp. 609–614.

[29] D S Johnson, A Mortazavi, R M Myers, et al. “Genome-wide mapping of in vivo protein-DNA interactions”. In: *Science* 316.5830 (2007), pp. 1497–1502. arXiv: 20.

[30] Tarjei S Mikkelsen, Manching Ku, David B Jaffe, et al. “Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.” In: *Nature* 448.7153 (2007), pp. 553–560.

[31] Ho Sung Rhee, Alain R. Bataille, Liye Zhang, et al. "Subnucleosomal structures and nucleosome asymmetry across a genome". In: *Cell* 159.6 (2014), pp. 1377–1388. arXiv: NIHMS150003.

[32] Peter J Skene and Steven Henikoff. "A simple method for generating high-resolution maps of genome-wide protein binding." In: *eLife* 4 (2015), e09225. arXiv: arXiv:1011.1669v3.

[33] Assaf Rotem, Oren Ram, Noam Shoresh, et al. "Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state". In: *Nature Biotechnology* 33.11 (2015), pp. 1165–72. arXiv: arXiv:0811.0484v1.

[34] Assaf Weiner, David Lara-Astiaso, Vladislav Krupalnik, et al. "Co-ChIP enables genome-wide mapping of histone mark co-occurrence at single-molecule resolution." In: *Nature biotechnology* 34.9 (2016), pp. 953–961.

[35] Raymond K. Auerbach, Ghia Euskirchen, Joel Rozowsky, et al. "Mapping accessible chromatin regions using Sono-Seq". In: *Proceedings of the National Academy of Sciences* 106.35 (2009), pp. 14926–14931.

[36] Monya Baker. "Blame it on the Antibodies". In: *Nature* 521.7552 (2015), pp. 274–275.

[37] Michele Busby, Catherine Xue, Catherine Li, et al. "Systematic comparison of monoclonal versus polyclonal antibodies for mapping histone modifications by ChIP-seq". In: *Epigenetics & Chromatin* 9.1 (2016), p. 49.

[38] Scott B. Rothbart, Bradley M. Dickson, Jesse R. Raab, et al. "An Interactive Database for the Assessment of Histone Antibody Specificity". In: *Molecular Cell* 59.3 (2015), pp. 502–511.

[39] D. Jain, S. Baldi, A. Zabel, et al. "Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments". In: *Nucleic Acids Research* 43.14 (2015), pp. 6959–68.

[40] Gregory E. Crawford, Ingeborg E. Holt, James Whittle, et al. "Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS)". In: *Genome Research* 16.1 (2006), pp. 123–131.

[41] Jay R Hesselberth, Xiaoyu Chen, Zhihong Zhang, et al. "Global mapping of protein-DNA interactions in vivo by digital genomic footprinting." In: *Nature methods* 6.4 (2009), pp. 283–9.

[42] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, et al. "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position." In: *Nature methods* 10.12 (2013), pp. 1213–8. arXiv: NIHMS150003.

[43] Hironori Waki, Masahiro Nakamura, Toshimasa Yamauchi, et al. "Global mapping of cell type-specific open chromatin by FAIRE-seq reveals the regulatory role of the NFI family in adipocyte differentiation". In: *PLoS Genetics* 7.10 (2011). Ed. by Jason D. Lieb, e1002311.

[44] Lingyun Song and Gregory E. Crawford. "DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells". In: *Cold Spring Harbor Protocols* 5.2 (2010), pdb.prot5384.

[45] J E Herrera and J B Chaires. *Characterization of preferred deoxyribonuclease I cleavage sites.* 1994.

[46] Pedro Madrigal. "On Accounting for Sequence-Specific Bias in Genome-Wide Chromatin Accessibility Experiments: Recent Advances and Contradictions". In: *Frontiers in Bioengineering and Biotechnology* 3.September (2015), pp. 1–4.

[47] Hashem Koohy, Thomas A. Down, and Tim J. Hubbard. "Chromatin Accessibility Data Sets Show Bias Due to Sequence Specificity of the DNase I Enzyme". In: *PLoS ONE* 8.7 (2013). Ed. by Leonardo Mariño-Ramírez, e69853.

[48] Galip Gürkan Yardimci, Christopher L. Frank, Gregory E. Crawford, et al. "Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection". In: *Nucleic Acids Research* 42.19 (2014), pp. 11865–11878.

[49] Eduardo G Gusmao, Manuel Allhoff, Martin Zenke, et al. "Analysis of computational footprinting methods for DNase sequencing experiments." In: *Nature methods* 13.4 (2016), pp. 303–9.

[50] Myong Hee Sung, Michael J. Guertin, Songjoon Baek, et al. "DNase footprint signatures are dictated by factor dynamics and DNA sequence". In: *Molecular Cell* 56.2 (2014), pp. 275–285. arXiv: NIHMS150003.

[51] Leighton J Core, Joshua J Waterfall, and John T Lis. "Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters." In: *Science (New York, N.Y.)* 322.5909 (2008), pp. 1845–8. arXiv: NIHMS150003.

[52] Hojoong Kwak, Nicholas J Fuda, Leighton J Core, et al. "Precise maps of RNA polymerase reveal how promoters direct initiation and pausing." In: *Science (New York, N.Y.)* 339.6122 (2013), pp. 950–3. arXiv: NIHMS150003.

[53] L S Churchman and J S Weissman. "Nascent transcript sequencing visualizes transcription at nucleotide resolution". In: *Nature* 469.7330 (2011), pp. 368–373.

[54] Joseph Rodriguez, Jerome S. Menet, and Michael Rosbash. "Nascent-Seq Indicates Widespread Cotranscriptional RNA Editing in Drosophila". In: *Molecular Cell* 47.1 (2012), pp. 27–37.

[55] B Schwalb, M Michel, B Zacher, et al. "TT-seq maps the human transient transcriptome". In: *Science* 352.6290 (2016), pp. 1225–1228.

[56] William S. Kruesi, Leighton J. Core, Colin T. Waters, et al. "Condensin controls recruitment of RNA polymerase ii to achieve nematode X-chromosome dosage compensation". In: *eLife* 2013.2 (2013), e00808.

[57] Ali Mortazavi, Brian A Williams, Kenneth McCue, et al. "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature Methods* 5.7 (2008), pp. 621–628. arXiv: 1111.6189v1.

[58] B Langmead, C Trapnell, M Pop, et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome". In: *Genome Biol.* 10.3 (2009), R25.

[59] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nat Methods* 9.4 (2012), pp. 357–359. arXiv: {\#}14603.

[60] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform". In: *Bioinformatics* 25.14 (2009), pp. 1754–1760.

[61] Daehwan Kim, Geo Pertea, Cole Trapnell, et al. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". In: *Genome Biology* 14.4 (2013), R36.

[62] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, et al. "STAR: Ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (2013), pp. 15–21.

[63] Bo Li and Colin N Dewey. "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." In: *BMC bioinformatics* 12.1 (2011), p. 323. arXiv: NIHMS150003.

[64] Alan P. Boyle, Justin Guinney, Gregory E. Crawford, et al. "F-Seq: A feature density estimator for high-throughput sequence tags". In: *Bioinformatics* 24.21 (2008), pp. 2537–2538.

[65] Yong Zhang, Tao Liu, Clifford A Meyer, et al. "Model-based Analysis of ChIP-Seq (MACS)". In: *Genome Biology* 9.9 (2008), R137.

[66] Naim U Rashid, Paul G Giresi, Joseph G Ibrahim, et al. "ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions". In: *Genome Biology* 12.7 (2011), R67.

[67] Qiang Song and Andrew D. Smith. "Identifying dispersed epigenomic domains from ChIP-Seq data". In: *Bioinformatics* 27.6 (2011), pp. 870–871.

[68] Haipeng Xing, Yifan Mo, Will Liao, et al. "Genome-wide localization of protein-DNA binding and histone modification by a bayesian change-point method with ChIP-seq data". In: *PLoS Computational Biology* 8.7 (2012), e1002613.

[69] Raja Jothi, Suresh Cuddapah, Artem Barski, et al. "Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data". In: *Nucleic Acids Research* 36.16 (2008), pp. 5221–5231.

[70] Anaïs F. Bardet, Jonas Steinmann, Sangeeta Bafna, et al. "Identification of transcription factor binding sites from ChIP-seq data at high resolution". In: *Bioinformatics* 29.21 (2013), pp. 2705–2713.

[71] Peter V Kharchenko, Michael Y Tolstorukov, and Peter J Park. "Design and analysis of ChIP-seq experiments for DNA-binding proteins." In: *Nature biotechnology* 26.12 (2008), pp. 1351–9.

[72] Vibhor Kumar, Masafumi Muratani, Nirmala Arul Rayan, et al. "Uniform, optimal signal processing of mapped deep-sequencing data". In: *Nature biotechnology* 31.7 (2013), pp. 1–11.

[73] Mahmoud M. Ibrahim, Scott A. Lacadie, and Uwe Ohler. "JAMM: A peak finder for joint analysis of NGS replicates". In: *Bioinformatics* 31.1 (2015), pp. 48–55.

[74] James H Bullard, Elizabeth Purdom, Kasper D Hansen, et al. "Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments". In: *U.C. Berkeley Div. Biostat. Pap. Ser.* 11.1 (2009), p. 94. arXiv: NIHMS150003.

[75] Marie Agnès Dillies, Andrea Rau, Julie Aubert, et al. "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis". In: *Briefings in Bioinformatics* 14.6 (2013), pp. 671–683.

[76] Mark Robinson and Alicia Oshlack. "A scaling normalization method for differential expression analysis of RNA-seq data". In: *Genome Biology* 11.3 (2010), R25.

[77] S Anders and W Huber. "Differential expression analysis for sequence count data". In: *Nature Precedings* 11.10 (2010), pp. 1–13. arXiv: 1310.0424.

[78] a Tsodikov, a Szabo, and D Jones. "Adjustments and measures of differential expression for microarray data." In: *Bioinformatics (Oxford, England)* 18.2 (2002), pp. 251–60.

[79] B. M. Bolstad, R. A. Irizarry, M. Åstrand, et al. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias". In: *Bioinformatics* 19.2 (2003), pp. 185–193.

[80] Charles Y. Lin, Jakob Lovén, Peter B. Rahl, et al. "Transcriptional amplification in tumor cells with elevated c-Myc". In: *Cell* 151.1 (2012), pp. 56–67.

[81] Zuqin Nie, Gangqing Hu, Gang Wei, et al. "c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells." In: *Cell* 151.1 (2012), pp. 68–79.

[82] Jakob Lovén, David A. Orlando, Alla A. Sigova, et al. "Revisiting global gene expression analysis". In: *Cell* 151.3 (2012), pp. 476–482.

[83] S. C. Hicks and R. A. Irizarry. "When to use Quantile Normalization?" In: *bioRxiv* (2014), p. 012203.

[84] Lichun Jiang, Felix Schlesinger, Carrie A. Davis, et al. "Synthetic spike-in standards for RNA-seq experiments". In: *Genome Research* 21.9 (2011), pp. 1543–1551.

[85] Nicolas Bonhoure, Gergana Bounova, David Bernasconi, et al. "Quantifying ChIP-seq data: A spiking method providing an internal reference for sample-to-sample normalization". In: *Genome Research* 24.7 (2014), pp. 1157–1168.

[86] David A. Orlando, Mei Wei Chen, Victoria E. Brown, et al. "Quantitative ChIP-Seq normalization reveals global modulation of the epigenome". In: *Cell Reports* 9.3 (2014), pp. 1163–1170.

[87] Adrian T. Grzybowski, Zhonglei Chen, and Alexander J. Ruthenburg. "Calibrating ChIP-Seq with Nucleosomal Internal Standards to Measure Histone Modification Density Genome Wide". In: *Molecular Cell* 58.5 (2015), pp. 886–899.

[88] M J Callow, S Dudoit, E L Gong, et al. "Microarray expression profiling identifies genes with altered expression in HDL-deficient mice." In: *Genome research* 10.12 (2000), pp. 2022–9.

[89] M K Kerr, C a Afshari, L Bennett, et al. "Statistical analysis of a gene expression microarray experiment with replication". In: *Statistica Sinica* 12.1 (2002), pp. 203–217.

[90] Xiangqin Cui and Gary a Churchill. "Statistical tests for differential expression in cDNA microarray experiments." In: *Genome biology* 4.4 (2003), p. 210.

[91] Xiangqin Gui, J. T Gene Hwang, Jing Qiu, et al. “Improved statistical tests for differential gene expression by shrinking variance components estimates”. In: *Biostatistics* 6.1 (2005), pp. 59–75.

[92] Mark D. Robinson and Gordon K. Smyth. “Moderated statistical tests for assessing differences in tag abundance”. In: *Bioinformatics* 23.21 (2007), pp. 2881–2887.

[93] F Rapaport, R Khanin, Y Liang, et al. “Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data.” In: *Genome biology* 14.9 (2013), R95. arXiv: arXiv:1301.5277v2.

[94] C Soneson and M Delorenzi. “A comparison of methods for differential expression analysis of RNA-seq data”. In: *BMC Bioinformatics* 14 (2013), p. 91.

[95] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. “edgeR: A Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26.1 (2009), pp. 139–140.

[96] Jaime Byrne, Patricia Nichols, Marzena Sroczynski, et al. “Prophylactic sacral dressing for pressure ulcer prevention in high-risk patients”. In: *American Journal of Critical Care* 25.3 (2016), pp. 228–234. arXiv: 1011.1669v3.

[97] Li Shen, Ning Yi Shao, Xiaochuan Liu, et al. “diffReps: Detecting Differential Chromatin Modification Sites from ChIP-seq Data with Biological Replicates”. In: *PLoS ONE* 8.6 (2013), e65598. arXiv: arXiv:1011.1669v3.

[98] Yanxiao Zhang, Yu Hsuan Lin, Timothy D. Johnson, et al. “PePr: A peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data”. In: *Bioinformatics* 30.18 (2014), pp. 2568–2575.

[99] Ming Yuan and Christina Kendziorski. “Hidden Markov Models for Microarray Time Course Data in Multiple Biological Conditions”. In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1323–1332.

[100] Ning Leng, Yuan Li, Brian E. McIntosh, et al. “EBSeq-HMM: A Bayesian approach for identifying gene-expression changes in ordered RNA-seq experiments”. In: *Bioinformatics* 31.16 (2015), pp. 2614–2622.

[101] X Wen, S Fuhrman, G S Michaels, et al. “Large-scale temporal gene expression mapping of central nervous system development.” In: *Proceedings of the National Academy of Sciences of the United States of America* 95.1 (1998), pp. 334–9.

[102] S Tavazoie, J D Hughes, M J Campbell, et al. “Systematic determination of genetic network architecture.” In: *Nature genetics* 22.3 (1999), pp. 281–5.

[103] M B Eisen, P T Spellman, P O Brown, et al. “Cluster analysis and display of genome-wide expression patterns.” In: *Proceedings of the National Academy of Sciences of the United States of America* 95.25 (1998), pp. 14863–8.

[104] Ziv Bar-Joseph, Erik D. Demaine, David K. Gifford, et al. “K-ary clustering with optimal leaf ordering for gene expression data”. In: *Bioinformatics* 19.9 (2003), pp. 1070–1078.

[105] P Tamayo, D Slonim, J Mesirov, et al. “Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.” In: *Proceedings of the National Academy of Sciences of the United States of America* 96.6 (1999), pp. 2907–2912.

[106] Michael P. S. Brown, William Noble Grundy, David Lin, et al. “Knowledge-based analysis of microarray gene expression data by using support vector machines”. In: *Proceedings of the National Academy of Sciences* 97.1 (2000), pp. 262–267.

[107] Ben-Dor Amir. “Clustering Gene Expression Patterns”. In: *Journal of Computational Biology* 6.3-4 (1999), pp. 281–297.

[108] K. Y. Yeung, C. Fraley, a. Murua, et al. “Model-based clustering and data transformations for gene expression data”. In: *Bioinformatics* 17.10 (2001), pp. 977–987.

[109] Z I V Bar-joseph, Georg K Gerber, David K Gifford, et al. “Continuous Representations of Gene Expression Data”. In: *Journal of Computational Biology* 10.3-4 (2003), pp. 341–356.

[110] Marco F Ramoni, Paola Sebastiani, and Isaac S Kohane. “Cluster analysis of gene expression dynamics”. In: *Pnas* 99.14 (2002), pp. 9121–9126.

[111] Alexander Schliep, Alexander Schönhuth, and Christine Steinhoff. “Using hidden Markov models to analyze gene expression time course data”. In: *Bioinformatics* 19.SUPPL. 1 (2003), pp. 255–263.

[112] Alexander Schliep, Christine Steinhoff, and Alexander Schönhuth. “Robust inference of groups in gene expression time-courses using mixtures of HMMs”. In: *Bioinformatics* 20.SUPPL. 1 (2004), pp. 283–289.

[113] Lue Ping Zhao, Ross Prentice, and Linda Breeden. “Statistical modeling of large microarray data sets to identify stimulus-response profiles”. In: *Proceedings of the National Academy of Sciences* 98.10 (2001), pp. 5631–5636.

[114] Y. Luan and H. Li. “Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data”. In: *Bioinformatics* 20.3 (2004), pp. 332–339.

[115] P D’haeseleer, X Wen, S Fuhrman, et al. “Linear modeling of mRNA expression levels during CNS development and injury.” In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 52 (1999), pp. 41–52.

[116] J. Aach and G. M. Church. “Aligning gene expression time series with time warping algorithms”. In: *Bioinformatics* 17.6 (2001), pp. 495–508.

[117] Jason Ernst, Gerard J. Nau, and Ziv Bar-Joseph. “Clustering short time series gene expression data”. In: *Bioinformatics* 21.SUPPL. 1 (2005), pp. 159–168.

[118] Jason Ernst, Ziv Bar-Joseph, PT Spellman, et al. “STEM: a tool for the analysis of short time series gene expression data”. In: *BMC Bioinformatics* 7.1 (2006), p. 191.

[119] Ziv Bar-Joseph. “Analyzing time series gene expression data”. In: *Bioinformatics* 20.16 (2004), pp. 2493–2503.

[120] P D’Haeseleer. “How does gene expression clustering work?” In: *Nature Biotechnology* 23.12 (2005), pp. 1499–1501.

[121] Ivan G Costa, Francisco De At de Carvalho, and Marcilio Cp de Souto. "Comparative analysis of clustering methods for gene expression time course data". In: *Genetics and Molecular Biology* 631.4 (2004), pp. 623–631.

[122] Sharon L. Paige, Sean Thomas, Cristi L. Stoick-Cooper, et al. "A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development". In: *Cell* 151.1 (2012), pp. 221–232. arXiv: NIHMS150003.

[123] Stefanie I. Becker. "Guidance of attention by feature relationships: The end of the road for feature map theories?" In: *Current Trends in Eye Tracking Research* 345.6199 (2014), pp. 37–49. arXiv: 1011.1669v3.

[124] Zheng Kuang, Ling Cai, Xuekui Zhang, et al. "High-temporal-resolution view of transcription and chromatin states across distinct metabolic states in budding yeast." In: *Nature structural & molecular biology* 21.10 (2014), pp. 854–63. arXiv: NIHMS150003.

[125] Pengfei Yu, Shu Xiao, Xiaoyun Xin, et al. "Spatiotemporal clustering of the epigenome reveals rules of dynamic gene regulation". In: *Genome Research* 23.2 (2013), pp. 352–364.

[126] N Friedman, M Linial, I Nachman, et al. "Using Bayesian networks to analyze expression data." In: *Journal of computational biology* 7.3-4 (2000), pp. 601–20.

[127] Mukesh Bansal, Vincenzo Belcastro, Alberto Ambesi-Impiombato, et al. "How to infer gene networks from expression profiles." In: *Molecular systems biology* 3.78 (2007), p. 78.

[128] Michael Hecker, Sandro Lambeck, Susanne Toepfer, et al. "Gene regulatory network inference: Data integration in dynamic models-A review". In: *BioSystems* 96.1 (2009), pp. 86–103.

[129] Lian En Chai, Swee Kuan Loh, Swee Thing Low, et al. "A review on the computational approaches for gene regulatory network construction". In: *Computers in Biology and Medicine* 48.1 (2014), pp. 55–65.

[130] Li-Hsieh Lin, Hsiao-Ching Lee, Wen-Hsiung Li, et al. "Dynamic modeling of cis-regulatory circuits and gene expression prediction via cross-gene identification." In: *BMC bioinformatics* 6.1 (2005), p. 258.

[131] Shawn Cokus, Sherri Rose, David Haynor, et al. "Modelling the network of cell cycle transcription factors in the yeast Saccharomyces cerevisiae." In: *BMC bioinformatics* 7.1 (2006), p. 381.

[132] Richard Bonneau, David J Reiss, Paul Shannon, et al. "The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo". In: *Genome Biology* 7.5 (2006), p. 1.

[133] Jason Ernst, Oded Vainas, Christopher T Harbison, et al. "Reconstructing dynamic regulatory maps." In: *Molecular systems biology* 3.74 (2007), p. 74.

[134] G P Georgiev. *Histones and the Control of Gene Action*. 1969.

[135] R C Huang and J Bonner. "Histone, a suppressor of chromosomal RNA synthesis." In: *Proceedings of the National Academy of Sciences of the United States of America* 48.7 (1962), pp. 1216–1222.

[136] Lawrence R. Gurley, J.Logan Irvin, and David J. Holbrook. “Inhibition of DNA polymerase by histones”. In: *Biochemical and Biophysical Research Communications* 14.6 (1964), pp. 527–532.

[137] H Weintraub and M Groudine. “Chromosomal subunits in active genes have an altered conformation.” In: *Science (New York, N.Y.)* 193.4256 (1976), pp. 848–56.

[138] Jamal Tazi and Adrian Bird. “Alternative chromatin structure at CpG islands”. In: *Cell* 60.6 (1990), pp. 909–920.

[139] Joseph A. Knezetic and Donal S. Luse. “The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro”. In: *Cell* 45.1 (1986), pp. 95–104.

[140] Swaminathan Venkatesh and Jerry L Workman. “Histone exchange, chromatin structure and the regulation of transcription.” In: *Nature Reviews. Molecular Cell Biology* 16.3 (2015), pp. 178–189.

[141] Ana Pombo and Niall Dillon. “Three-dimensional genome architecture: players and mechanisms.” In: *Nature reviews. Molecular cell biology* 16.4 (2015), pp. 245–257.

[142] Varun Narendra, Pedro P Rocha, Disi An, et al. “Transcription. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation.” In: *Science (New York, N.Y.)* 347.6225 (2015), pp. 1017–21.

[143] Matthias Merkenschlager and Duncan T. Odom. “CTCF and cohesin: Linking gene regulatory elements with their targets”. In: *Cell* 152.6 (2013), pp. 1285–1297.

[144] Michael E. Dobson and Vernon M. Ingram. “In vitro transcription of chromatin containing histones hyperacetylated in vivo”. In: *Nucleic Acids Research* 8.18 (1980), pp. 4201–4220.

[145] Diane J. Mathis, Pierre Oudet, Bohdan Wasylyk, et al. “Effect of histone acetylation on structure and in vitro transcription of chromatin”. In: *Nucleic Acids Research* 5.10 (1978), pp. 3523–3548.

[146] He Huang, Benjamin R. Sabari, Benjamin A. Garcia, et al. “SnapShot: Histone modifications”. In: *Cell* 159.2 (2014), 458–458.e1.

[147] P. N. I. Lau and P. Cheung. “Histone code pathway involving H3 S28 phosphorylation and K27 acetylation activates transcription and antagonizes polycomb silencing”. In: *Proceedings of the National Academy of Sciences* 108.7 (2011), pp. 2801–2806.

[148] C Tse, T Sera, a P Wolffe, et al. “Disruption of higher-order folding by core histone acetylation dramatically enhances transcription of nucleosomal arrays by RNA polymerase III.” In: *Molecular and cellular biology* 18.8 (1998), pp. 4629–4638.

[149] T Zhang, S Cooper, and N Brockdorff. “The interplay of histone modifications - writers that read”. In: *EMBO Rep* 16.11 (2015), pp. 1467–1481.

[150] Vasily V. Ogryzko, R. Louis Schiltz, Valya Russanova, et al. “The transcriptional coactivators p300 and CBP are histone acetyltransferases”. In: *Cell* 87.5 (1996), pp. 953–959.

[151] Dirk Schübeler. “Function and information content of DNA methylation”. In: *Nature* 517.7534 (2015), pp. 321–326.

[152] Moyra Lawrence, Sylvain Daujat, and Robert Schneider. "Lateral Thinking: How Histone Modifications Regulate Gene Expression". In: *Trends in Genetics* 32.1 (2016), pp. 42–56.

[153] Miler T Lee, Ashley R Bonneau, and Antonio J Giraldez. "Zygotic Genome Activation During the Maternal-to- Zygotic Transition". In: *Annu. Rev. Cell Dev. Biol* 30.1 (2014), pp. 581–613. arXiv: NIHMS150003.

[154] Sarita S. Paranjpe and Gert Jan C Veenstra. "Establishing pluripotency in early development". In: *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* 1849.6 (2015), pp. 626–636.

[155] Silvia Velasco, Mahmoud M Ibrahim, Akshay Kakumanu, et al. "A multi-step transcriptional and chromatin cascade underlies motor neuron programming". In: *Cell Stem Cell* In press (2016).

[156] H. J. Rippon and A. E. Bishop. *Embryonic stem cells*. 2004.

[157] Gordon Keller. "Embryonic stem cell differentiation : emergence of a new era in biology and medicine". In: *Genes & development* 19 (2005), pp. 1129–1155.

[158] Allen Wang, Feng Yue, Yan Li, et al. "Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates". In: *Cell Stem Cell* 16.4 (2015), pp. 386–399.

[159] Patrick Cahan, Hu Li, Samantha A. Morris, et al. "CellNet: Network biology applied to stem cell engineering". In: *Cell* 158.4 (2014), pp. 903–915.

[160] Esteban O Mazzoni, Shaun Mahony, Michael Closser, et al. "Synergistic binding of transcription factors to cell-specific enhancers programs motor neuron identity." In: *Nature neuroscience* 16.9 (2013), pp. 1219–27.

[161] Hiroshi Kurosawa. "Methods for inducing embryoid body formation: in vitro differentiation system of embryonic stem cells". In: *Journal of Bioscience and Bioengineering* 103.5 (2007), pp. 389–398.

[162] Julia Ladewig, Philipp Koch, and Oliver Brüstle. "Leveling Waddington: the emergence of direct programming and the loss of cell fate hierarchies." In: *Nature reviews. Molecular cell biology* 14.4 (2013), pp. 225–36.

[163] Jun Xu, Yuanyuan Du, and Hongkui Deng. "Direct lineage reprogramming: Strategies, mechanisms, and applications". In: *Cell Stem Cell* 16.2 (2015), pp. 119–134.

[164] Kazutoshi Takahashi and Shinya Yamanaka. "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors". In: *Cell* 126.4 (2006), pp. 663–676.

[165] Thomas Vierbuchen, Austin Ostermeier, Zhiping P. Pang, et al. "Direct conversion of fibroblasts to functional neurons by defined factors". In: *Nature* 463.7284 (2010), pp. 1035–1041.

[166] Abdenour Soufi, Greg Donahue, and Kenneth S. Zaret. "Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome". In: *Cell* 151.5 (2012), pp. 994–1004. arXiv: NIHMS150003.

[167] Barbara Treutlein, Qian Yi Lee, J. Gray Camp, et al. “Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq”. In: *Nature* 534.7607 (2016), p. 391395.

[168] Tamar Juven-Gershon, Jer Yuan Hsu, Joshua W. Theisen, et al. “The RNA polymerase II core promoter - the gateway to transcription”. In: *Current Opinion in Cell Biology* 20.3 (2008), pp. 253–259.

[169] Tamar Juven-Gershon and James T. Kadonaga. “Regulation of gene expression via the core promoter and the basal transcriptional machinery”. In: *Developmental Biology* 339.2 (2010), pp. 225–229.

[170] Manju Bansal, Aditya Kumar, and Venkata Rajesh Yella. “Role of DNA sequence based structural features of promoters in transcription initiation and gene expression”. In: *Current Opinion in Structural Biology* 25 (2014), pp. 77–85.

[171] Uwe Ohler. “Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction”. In: *Nucleic Acids Research* 34.20 (2006), pp. 5943–5950.

[172] Sascha H C Duttke. “Evolution and diversification of the basal transcription machinery”. In: *Trends in Biochemical Sciences* 40.3 (2015), pp. 127–129.

[173] Simon J. Van Heeringen, Waseem Akhtar, Ulrike G. Jacobi, et al. “Nucleotide composition-linked divergence of vertebrate core promoter architecture”. In: *Genome Research* 21.3 (2011), pp. 410–421.

[174] J. E F Butler and J. T. Kadonaga. “Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs”. In: *Genes and Development* 15.19 (2001), pp. 2515–2519.

[175] S. T. Smale. “Core promoters: Active contributors to combinatorial gene regulation”. In: *Genes and Development* 15.19 (2001), pp. 2503–2508.

[176] Kevin J Wright, Michael T Marr, and Robert Tjian. “TAF4 nucleates a core subcomplex of TFIID and mediates activated transcription from a TATA-less promoter.” In: *Proceedings of the National Academy of Sciences of the United States of America* 103.33 (2006), pp. 12347–52.

[177] James A. Goodrich and Robert Tjian. “Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation”. In: *Nature Reviews Genetics* 11.8 (2010), pp. 549–558.

[178] C. Peter Verrijzer, Jin Long Chen, Kyoko Yokomori, et al. “Binding of TAFs to core elements directs promoter selectivity by RNA polymerase II”. In: *Cell* 81.7 (1995), pp. 1115–1125.

[179] P Carninci, A Sandelin, B Lenhard, et al. “Genome-wide analysis of mammalian promoter architecture and evolution”. In: *Nature genetics* 38.6 (2006), pp. 626–635.

[180] José M. Morachis, Christopher M. Murawsky, and Beverly M. Emerson. “Regulation of the p53 transcriptional response by structurally diverse core promoters”. In: *Genes and Development* 24.2 (2010), pp. 135–147.

[181] Andreas Hochheimer and Robert Tjian. “Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression”. In: *Genes and Development* 17.11 (2003), pp. 1309–1320.

[182] Daniel Zenklusen, Daniel R Larson, and Robert H Singer. “Single-RNA counting reveals alternative modes of gene expression in yeast. TL - 15”. In: *Nature structural & molecular biology* 15 VN - r.12 (2008), pp. 1263–1271.

[183] Keren Bahar Halpern, Sivan Tanami, Shanie Landen, et al. “Bursty gene expression in the intact mammalian liver”. In: *Molecular Cell* 58.1 (2015), pp. 147–156.

[184] Arjun Raj, Charles S. Peskin, Daniel Tranchina, et al. “Stochastic mRNA synthesis in mammalian cells”. In: *PLoS Biology* 4.10 (2006). Ed. by Ueli Schibler, pp. 1707–1719.

[185] Jonathan R. Chubb, Tatjana Trcek, Shailesh M. Shenoy, et al. “Transcriptional Pulsing of a Developmental Gene”. In: *Current Biology* 16.10 (2006), pp. 1018–1025.

[186] John R S Newman, Sina Ghaemmaghami, Jan Ihmels, et al. “Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise TL - 441”. In: *Nature* 441 VN -.7095 (2006), pp. 840–846.

[187] Kathryn Miller-Jensen, Siddharth S. Dey, David V. Schaffer, et al. “Varying virulence: Epigenetic control of expression noise and disease processes”. In: *Trends in Biotechnology* 29.10 (2011), pp. 517–525.

[188] Hinrich Boeger, Joachim Griesenbeck, and Roger D. Kornberg. “Nucleosome Retention and the Stochastic Nature of Promoter Chromatin Remodeling for Transcription”. In: *Cell* 133.4 (2008), pp. 716–726.

[189] David M Suter, Nacho Molina, David Gatfield, et al. “with Widely Different Bursting Kinetics”. In: *Science* 332.iii (2011), pp. 21–24.

[190] Ferenc Müller and Làszlò Tora. “Chromatin and DNA sequences in defining promoters for transcription initiation”. In: *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* 1839.3 (2014), pp. 118–128.

[191] Robert E Thurman, Eric Rynes, Richard Humbert, et al. “The accessible chromatin landscape of the human genome.” In: *Nature* 489.7414 (2012), pp. 75–82. arXiv: NIHMS150003.

[192] Travis N Mavrich, Cizhong Jiang, Ilya P Ioshikhes, et al. “Nucleosome organization in the Drosophila genome.” In: *Nature* 453.7193 (2008), pp. 358–362.

[193] J W Ho, Y L Jung, T Liu, et al. “Comparative analysis of metazoan chromatin organization”. In: *Nature* 512.7515 (2014), pp. 449–452. arXiv: NIHMS150003.

[194] Kevin Struhl and Eran Segal. “Determinants of nucleosome positioning.” In: *Nature structural & molecular biology* 20.3 (2013), pp. 267–73.

[195] Eivind Valen and Albin Sandelin. “Genomic and chromatin signals underlying transcription start-site selection”. In: *Trends in Genetics* 27.11 (2011), pp. 475–485. arXiv: 0002067 [hep-th].

[196] Paul D. Hartley and Hiten D. Madhani. “Mechanisms that Specify Promoter Nucleosome Location and Identity”. In: *Cell* 137.3 (2009), pp. 445–458.

[197] Itay Tirosh, Naama Barkai, and Kevin J Verstrepen. “Promoter architecture and the evolvability of gene expression”. In: *Journal of biology* 8.11 (2009), p. 95.

[198] Lori L. Wallrath, Quinn Lu, Howard Granok, et al. “Architectural variations of inducible eukaryotic promoters: Preset and remodeling chromatin structures”. In: *BioEssays* 16.3 (1994), pp. 165–170.

[199] Ting Ni, David L Corcoran, Elizabeth A Rach, et al. “A paired-end sequencing strategy to map the complex landscape of transcription initiation.” In: *Nature methods* 7.7 (2010), pp. 521–7.

[200] Roger A. Hoskins, Jane M. Landolin, James B. Brown, et al. “Genome-wide analysis of promoter architecture in Drosophila melanogaster”. In: *Genome Research* 21.2 (2011), pp. 182–192.

[201] Nicolas Nègre, Christopher D. Brown, Parantu K. Shah, et al. “A comprehensive map of insulator elements for the Drosophila genome”. In: *PLoS Genetics* 6.1 (2010). Ed. by Yoshihide Hayashizaki, e1000814.

[202] Elizabeth A. Rach, Deborah R. Winter, Ashlee M. Benjamin, et al. “Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level”. In: *PLoS Genetics* 7.1 (2011). Ed. by Jason D. Lieb, e1001274.

[203] Itay Tirosh and Naama Barkai. “Two strategies for gene regulation by promoter nucleosomes”. In: *Genome Research* 18.7 (2008), pp. 1084–1091.

[204] Stavros Lomvardas and Dimitris Thanos. “Modifying gene expression programs by altering core promoter chromatin architecture”. In: *Cell* 110.2 (2002), pp. 261–271.

[205] Vladimir R. Ramirez-Carrozzi, Daniel Braas, Dev M. Bhatt, et al. “A Unifying Model for the Selective Regulation of Inducible Transcription by CpG Islands and Nucleosome Remodeling”. In: *Cell* 138.1 (2009), pp. 114–128.

[206] Bradley R Cairns. “The logic of chromatin architecture and remodelling at promoters.” In: *Nature* 461.7261 (2009), pp. 193–198.

[207] Elizabeth A Rach, Hsiang-Yu Yuan, William H Majoros, et al. “Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome.” In: *Genome biology* 10.7 (2009), R73.

[208] Nathaniel D Heintzman, Rhona K Stuart, Gary Hon, et al. “Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.” In: *Nature genetics* 39.3 (2007), pp. 311–8.

[209] Artem Barski, Suresh Cuddapah, Kairong Cui, et al. “High-Resolution Profiling of Histone Methylations in the Human Genome”. In: *Cell* 129.4 (2007), pp. 823–837.

[210] Antonin Morillon, Nickoletta Karabetsou, Anitha Nair, et al. “Dynamic lysine methylation on histone H3 defines the regulatory phase of gene transcription”. In: *Molecular Cell* 18.6 (2005), pp. 723–734.

[211] Zhibin Wang, Chongzhi Zang, Jeffrey A Rosenfeld, et al. “Combinatorial patterns of histone acetylations and methylations in the human genome.” In: *Nature genetics* 40.7 (2008), pp. 897–903.

[212] Peter V Kharchenko, Artyom A Alekseyenko, Yuri B Schwartz, et al. "Comprehensive analysis of the chromatin landscape in Drosophila melanogaster." In: *Nature* 471.7339 (2011), pp. 480–5.

[213] Ali Shilatifard. "The COMPASS Family of Histone H3K4 Methylases: Mechanisms of Regulation in Development and Disease Pathogenesis". en. In: *Biochemistry* 81.1 (2012), pp. 65–95.

[214] Michiel Vermeulen, Klaas W. Mulder, Sergei Denissov, et al. "Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4". English. In: *Cell* 131.1 (2007), pp. 58–69.

[215] Shannon M. Lauberth, Takahiro Nakayama, Xiaolin Wu, et al. "H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation". In: *Cell* 152.5 (2013), pp. 1021–1036. arXiv: NIHMS150003.

[216] Craig A. Mizzen, Xiang Jiao Yang, Tetsuro Kokubo, et al. "The TAF(II)250 subunit of TFIID has histone acetyltransferase activity". In: *Cell* 87.7 (1996), pp. 1261–1270.

[217] Timothy J. Stasevich, Yoko Hayashi-Takanaka, Yuko Sato, et al. "Regulation of RNA polymerase II activation by histone acetylation in single living cells." In: *Nature* 516.7530 (2014), pp. 272–5.

[218] Rosa Karlić, Ho-Ryun Chung, Julia Lasserre, et al. "Histone modification levels are predictive for gene expression". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.7 (2010), pp. 2926–2931.

[219] Yun Chen, Mette Jorgensen, Raivo Kolde, et al. "Prediction of RNA Polymerase II recruitment, elongation and stalling from histone modification data." In: *BMC genomics* 12.1 (2011), p. 544.

[220] Ramana V. Davuluri, Yutaka Suzuki, Sumio Sugano, et al. "The functional consequences of alternative promoter use in mammalian genomes". In: *Trends in Genetics* 24.4 (2008), pp. 167–177.

[221] Philippe Batut, Alexander Dobin, Charles Plessy, et al. "High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression". In: *Genome Research* 23.1 (2013), pp. 169–180. arXiv: 15334406.

[222] Yong Zhang, Nadine L. Vastenhouw, Jianxing Feng, et al. "Canonical nucleosome organization at promoters forms during genome activation". In: *Genome Research* 24.2 (2014), pp. 260–266.

[223] Chirag Nepal, Yavor Hadzhiev, Christopher Previti, et al. "Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis". In: *Genome Research* 23.11 (2013), pp. 1938–1950.

[224] Tetsuya Muramoto, Danielle Cannon, Marek Gierlinski, et al. "Live imaging of nascent RNA dynamics reveals distinct types of transcriptional pulse regulation." In: *Proceedings of the National Academy of Sciences of the United States of America* 109.19 (2012), pp. 7350–7355.

[225] M Prioleau. "Competition between chromatin and transcription complex assembly regulates gene expression during early development". In: *Cell* 77.3 (1994), pp. 439–449.

[226] Philipp Voigt, Wee Wei Tee, and Danny Reinberg. “A double take on bivalent promoters”. In: *Genes and Development* 27.12 (2013), pp. 1318–1338.

[227] Tobias D. Schneider, Jose M. Arteaga-Salas, Edith Mentele, et al. “Stage-Specific histone modification profiles reveal global transitions in the Xenopus embryonic epigenome”. In: *PLoS ONE* 6.7 (2011). Ed. by Ferenc Mueller, e22548.

[228] Keisuke Aoshima, Erina Inoue, Hirofumi Sawa, et al. “Paternal H3K4 methylation is required for minor zygotic gene activation and early mouse embryonic development.” In: *EMBO reports* 16.7 (2015), pp. 803–12.

[229] Peter J Rugg-Gunn, Brian J Cox, Amy Ralston, et al. “Distinct histone modifications in stem cell lines and tissue lineages from the early mouse embryo.” In: *Proceedings of the National Academy of Sciences of the United States of America* 107.24 (2010), pp. 10783–90.

[230] Bradley E. Bernstein, Tarjei S. Mikkelsen, Xiaohui Xie, et al. “A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells”. In: *Cell* 125.2 (2006), pp. 315–326.

[231] Véronique Azuara, Pascale Perry, Stephan Sauer, et al. “Chromatin signatures of pluripotent cell lines.” In: *Nature cell biology* 8.5 (2006), pp. 532–8.

[232] Philipp Voigt, Gary LeRoy, William J. Drury, et al. “Asymmetrically modified nucleosomes”. In: *Cell* 151.1 (2012), pp. 181–193.

[233] Guangjin Pan, Shulan Tian, Jeff Nie, et al. “Whole-Genome Analysis of Histone H3 Lysine 4 and Lysine 27 Methylation in Human Embryonic Stem Cells”. In: *Cell Stem Cell* 1.3 (2007), pp. 299–312.

[234] Xiao Dong Zhao, Xu Han, Joon Lin Chew, et al. “Whole-Genome Mapping of Histone H3 Lys4 and 27 Trimethylations Reveals Distinct Genomic Compartments in Human Embryonic Stem Cells”. In: *Cell Stem Cell* 1.3 (2007), pp. 286–298.

[235] Wei Xie, Matthew D. Schultz, Ryan Lister, et al. “Epigenomic analysis of multilineage differentiation of human embryonic stem cells”. In: *Cell* 153.5 (2013), pp. 1134–1148.

[236] Eric M. Mendenhall, Richard P. Koche, Thanh Truong, et al. “GC-rich sequence elements recruit PRC2 in mammalian ES cells”. In: *PLoS Genetics* 6.12 (2010). Ed. by Hiten D. Madhani, pp. 1–10.

[237] Tong Ihn Lee, Richard G. Jenner, Laurie A. Boyer, et al. “Control of Developmental Regulators by Polycomb in Human Embryonic Stem Cells”. In: *Cell* 125.2 (2006), pp. 301–313. arXiv: NIHMS150003.

[238] Ru Cao, L Wang, Hengbin Wang, et al. “Role of histone H3 lysine 27 methylation in Polycomb group silencing”. In: *Science* 298.November (2002), pp. 1039–1044.

[239] Liangjun Wang, J. Lesley Brown, Ru Cao, et al. “Hierarchical recruitment of polycomb group silencing complexes”. In: *Molecular Cell* 14.5 (2004), pp. 637–646.

[240] Lynn Lehmann, Roberto Ferrari, Ajay A. Vashisht, et al. “Polycomb repressive complex 1 (PRC1) disassembles RNA polymerase II preinitiation complexes”. In: *Journal of Biological Chemistry* 287.43 (2012), pp. 35784–35794.

[241] Julie K Stock, Sara Giadrossi, Miguel Casanova, et al. "Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells." In: *Nature cell biology* 9.12 (2007), pp. 1428–1435.

[242] Wenlai Zhou, Ping Zhu, Jianxun Wang, et al. "Histone H2A Monoubiquitination Represses Transcription by Inhibiting RNA Polymerase II Transcriptional Elongation". In: *Molecular Cell* 29.1 (2008), pp. 69–80. arXiv: NIHMS150003.

[243] Irene M. Min, Joshua J. Waterfall, Leighton J. Core, et al. "Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells". In: *Genes and Development* 25.7 (2011), pp. 742–754.

[244] Emily Brookes, Inês De Santiago, Daniel Hebenstreit, et al. "Polycomb associates genome-wide with a specific RNA polymerase II variant, and regulates metabolic genes in ESCs". In: *Cell Stem Cell* 10.2 (2012), pp. 157–170.

[245] Aditi Kanhere, Keijo Viiri, Carla C. Araújo, et al. "Short RNAs Are Transcribed from Repressed Polycomb Target Genes and Interact with Polycomb Repressive Complex-2". In: *Molecular Cell* 38.5 (2010), pp. 675–688. arXiv: Kanhere20542000,.

[246] Emily Walker, Wing Y. Chang, Julie Hunkapiller, et al. "Polycomb-like 2 Associates with PRC2 and Regulates Transcriptional Networks during Mouse Embryonic Stem Cell Self-Renewal and Differentiation". In: *Cell Stem Cell* 6.2 (2010), pp. 153–166.

[247] Laurie A Boyer, Kathrin Plath, Julia Zeitlinger, et al. "Polycomb complexes repress developmental regulators in murine embryonic stem cells." In: *Nature* 441.7091 (2006), pp. 349–353.

[248] Sergei Denissov, Helmut Hofemeister, Hendrik Marks, et al. "Mll2 is required for H3K4 trimethylation on bivalent promoters in embryonic stem cells, whereas Mll1 is redundant." In: *Development (Cambridge, England)* 141.3 (2014), pp. 526–37.

[249] Frank W. Schmitges, Archana B. Prusty, Mahamadou Faty, et al. "Histone Methylation by PRC2 Is Inhibited by Active Chromatin Marks". In: *Molecular Cell* 42.3 (2011), pp. 330–341.

[250] Diego Pasini, Martina Malatesta, Hye Ryung Jung, et al. "Characterization of an antagonistic switch between histone H3 lysine 27 methylation and acetylation in the transcriptional regulation of Polycomb group target genes". In: *Nucleic Acids Research* 38.15 (2010), pp. 4958–4969.

[251] R David Hawkins, Gary C Hon, Chuhu Yang, et al. "Dynamic chromatin states in human ES cells reveal potential regulatory sequences and genes involved in pluripotency". In: *Cell Research* 21.10 (2011), pp. 1393–1409.

[252] Feng Tie, Rakhee Banerjee, Alina R Saiakhova, et al. "Trithorax monomethylates histone H3K4 and interacts directly with CBP to promote H3K27 acetylation and antagonize Polycomb silencing." In: *Development (Cambridge, England)* 141 (2014), pp. 1129–39.

[253] Feng Tie, Rakhee Banerjee, Chen Fu, et al. "Polycomb inhibits histone acetylation by CBP by binding directly to its catalytic domain." In: *Proceedings of the National Academy of Sciences of the United States of America* 113.6 (2016), E744–53.

[254] Felix Muerdter and Alexander Stark. “Gene Regulation: Activation through Space”. In: *Current Biology* 26.19 (2016), R895–R898.

[255] Tae Kyung Kim and Ramin Shiekhattar. “Architectural and Functional Commonalities between Enhancers and Promoters”. In: *Cell* 162.5 (2015), pp. 948–959. arXiv: NIHMS150003.

[256] Takashi Fukaya, Bomyi Lim, and Michael Levine. “Enhancer Control of Transcriptional Bursting”. In: *Cell* 166.2 (2016), pp. 358–368.

[257] Caroline R. Bartman, Sarah C. Hsu, Chris C S Hsiung, et al. “Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping”. In: *Molecular Cell* 62.2 (2016), pp. 237–247.

[258] Katjana Tantale, Florian Mueller, Alja Kozulic-Pirher, et al. “A single-molecule view of transcription reveals convoys of RNA polymerases and multi-scale bursting.” In: *Nature communications* 7 (2016), p. 12248.

[259] Falong Lu, Yuting Liu, Azusa Inoue, et al. “Establishing chromatin regulatory landscape during mouse preimplantation development”. In: *Cell* 165.6 (2016), pp. 1375–1388.

[260] Andrew B. Stergachis, Shane Neph, Alex Reynolds, et al. “Developmental fate and cellular maturity encoded in human regulatory DNA landscapes”. In: *Cell* 154.4 (2013), pp. 888–903. arXiv: NIHMS150003.

[261] Jingyi Wu, Bo Huang, He Chen, et al. “The landscape of accessible chromatin in mammalian preimplantation embryos.” In: *Nature* 534.7609 (2016), pp. 652–7.

[262] Kashif Ahmed, Hesam Dehghani, Peter Rugg-Gunn, et al. “Global chromatin architecture reflects pluripotency and lineage commitment in the early mouse embryo”. In: *PLoS ONE* 5.5 (2010). Ed. by Axel Imhof, e10531.

[263] Ana Bošković, André Eid, Julien Pontabry, et al. “Higher chromatin mobility supports totipotency and precedes pluripotency in vivo”. In: *Genes and Development* 28.10 (2014), pp. 1042–1047.

[264] Bo Wen, Hao Wu, Yoichi Shinkai, et al. “Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells.” In: *Nature genetics* 41.2 (2009), pp. 246–50.

[265] Tadaatsu Goto, Paul Macdonald, and Tom Maniatis. “Early and late periodic patterns of even skipped expression are controlled by distinct regulatory elements that respond to different spatial cues”. In: *Cell* 57.3 (1989), pp. 413–422.

[266] K Harding, T Hoey, R Warrior, et al. “Autoregulatory and gap gene response elements of the even-skipped promoter of Drosophila.” In: *The EMBO journal* 8.4 (1989), pp. 1205–12.

[267] Sheldon Rowan, Trevor Siggers, Salil A. Lachke, et al. “Precise temporal control of the eye regulatory gene Pax6 via enhancer-binding site affinity”. In: *Genes and Development* 24.10 (2010), pp. 980–985.

[268] Kiwon Lee, Chris C Hsiung, Peng Huang, et al. “Dynamic enhancer Gene body contacts during transcription elongation”. In: *Genes & Development* (2015), pp. 1992–1997.

[269] C T Ong and V G Corces. “Enhancers: emerging roles in cell fate specification”. In: *EMBO Rep* 13.5 (2012), pp. 423–430.

[270] M. Ptashne, A. Jeffrey, A. D. Johnson, et al. “How the lambda repressor and cro work”. In: *Cell* 19.1 (1980), pp. 1–11.

[271] Joana Osório. “Landscape and mechanisms of transcription factor cooperativity”. In: *Nature Publishing Group* 15545.November (2015), p. 15545.

[272] Outi Hallikas, Kimmo Palin, Natalia Sinjushina, et al. “Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity”. In: *Cell* 124.1 (2006), pp. 47–59.

[273] Chaolin Zhang, Zhenyu Xuan, Stefanie Otto, et al. “A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome”. In: *Nucleic Acids Research* 34.8 (2006), pp. 2238–2246.

[274] Arttu Jolma, Yimeng Yin, Kazuhiro R. Nitta, et al. “DNA-dependent formation of transcription factor pairs alters their binding specificity”. In: *Nature* 527.7578 (2015), pp. 384–8.

[275] Enrico Cannavò, Pierre Khoueiry, David A. Garfield, et al. “Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks”. In: *Current Biology* 26.1 (2016), pp. 38–51.

[276] Deborah Hay, Jim R Hughes, Christian Babbs, et al. “Genetic dissection of the α-globin super-enhancer in vivo.” In: *Nature genetics* 48.July (2016), pp. 1–12.

[277] Edwin Smith and Ali Shilatifard. “Enhancer biology and enhanceropathies.” In: *Nature structural & molecular biology* 21.3 (2014), pp. 210–9.

[278] Joung-Woo Hong, David A Hendrix, and Michael S Levine. “Shadow Enhancers as a Source of Evolutionary Novelty”. In: *Science* 321.5894 (2008), p. 1314.

[279] Winship Herr and Jennifer Clarke. “The SV40 enhancer is composed of multiple functional elements that can compensate for one another”. In: *Cell* 45.3 (1986), pp. 461–470.

[280] Kenneth S. Zaret and Jason S. Carroll. “Pioneer transcription factors: Establishing competence for gene expression”. In: *Genes and Development* 25.21 (2011), pp. 2227–2241.

[281] Abdenour Soufi, Meilin Fernandez Garcia, Artur Jaroszewicz, et al. “Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming.” In: *Cell* 161.3 (2015), pp. 555–68.

[282] Bradley E. Bernstein, Michael Kamal, Kerstin Lindblad-Toh, et al. “Genomic maps and comparative analysis of histone modifications in human and mouse”. In: *Cell* 120.2 (2005), pp. 169–181.

[283] Nathaniel D Heintzman, Gary C Hon, R David Hawkins, et al. “Histone modifications at human enhancers reflect global cell-type-specific gene expression.” In: *Nature* 459.7243 (2009), pp. 108–112.

[284] Sven Heinz, Christopher Benner, Nathanael Spann, et al. “Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities”. In: *Molecular Cell* 38.4 (2010), pp. 576–589.

[285] Menno P Creyghton, Albert W Cheng, G Grant Welstead, et al. “Histone H3K27ac separates active from poised enhancers and predicts developmental state.” In: *Proceedings of the National Academy of Sciences of the United States of America* 107.50 (2010), pp. 21931–21936. arXiv: arXiv:1408.1149.

[286] Alvaro Rada-Iglesias, Ruchi Bajpai, Tomek Swigut, et al. “A unique chromatin signature uncovers early developmental enhancers in humans.” In: *Nature* 470.7333 (2011), pp. 279–83. arXiv: 15334406.

[287] Minna U. Kaikkonen, Nathanael J. Spann, Sven Heinz, et al. “Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription”. In: *Molecular Cell* 51.3 (2013), pp. 310–325.

[288] Axel Visel, Matthew J Blow, Zirong Li, et al. “ChIP-seq accurately predicts tissue-specific activity of enhancers.” In: *Nature* 457.7231 (2009), pp. 854–8.

[289] Tae-Kyung Kim, Martin Hemberg, Jesse M Gray, et al. *Widespread transcription at neuronal activity-regulated enhancers.* 2010.

[290] Francesca de Santa, Iros Barozzi, Flore Mietton, et al. “A large fraction of extragenic RNA Pol II transcription sites overlap enhancers”. In: *PLoS Biology* 8.5 (2010). Ed. by John S. Mattick, e1000384.

[291] Dong Wang, Ivan Garcia-Bassets, Chris Benner, et al. “Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA.” In: *Nature* 474.7351 (2011), pp. 390–4.

[292] Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, et al. “An atlas of active enhancers across human cell types and tissues.” In: *Nature* 507.7493 (2014), pp. 455–61.

[293] Robert S Young, Yatendra Kumar, Wendy A Bickmore, et al. “Bidirectional transcription marks accessible chromatin and is not specific to enhancers”. In: *bioRxiv* (2016), p. 048629.

[294] Cosmas D. Arnold, Daniel Gerlach, Christoph Stelzer, et al. “Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq”. In: *Science* 339.6123 (2013), pp. 1074–1077.

[295] Stephen S Gisselbrecht, Luis A Barrera, Martin Porsch, et al. “Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos.” In: *Nature methods* 10.8 (2013), pp. 774–80. arXiv: NIHMS150003.

[296] Jamie C. Kwasnieski, Christopher Fiore, Hemangi G. Chaudhari, et al. “High-throughput functional testing of ENCODE segmentation predictions”. In: *Genome Research* 24.10 (2014), pp. 1595–1602.

[297] Matthew Murtha, Zeynep Tokcaer-Keskin, Zuojian Tang, et al. “FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells.” In: *Nature methods* 11.5 (2014), pp. 559–65. arXiv: NIHMS150003.

[298] Sascha H C Duttke, Scott A. Lacadie, Mahmoud M. Ibrahim, et al. “Human promoters are intrinsically directional”. In: *Molecular Cell* 57.4 (2015), pp. 674–684.

[299] Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, et al. "Mapping and analysis of chromatin state dynamics in nine human cell types". In: *Nature* 473.7345 (2011), pp. 43–9.

[300] Yin Shen, Feng Yue, David F. McCleary, et al. "A map of the cis-regulatory sequences in the mouse genome". In: *Nature* 488.7409 (2012), pp. 116–120. arXiv: NIHMS150003.

[301] D. Lara-Astiaso, A. Weiner, E. Lorenzo-Vivas, et al. "Chromatin state dynamics during blood formation". In: *Science* 345.6199 (2014), pp. 943–9. arXiv: arXiv:1011.1669v3.

[302] Anirudh Natarajan, Galip Gürkan Yardimci, Nathan C. Sheffield, et al. "Predicting cell-type-specific gene expression from regions of open chromatin". In: *Genome Research* 22.9 (2012), pp. 1711–1722.

[303] Weronika Sikora-Wohlfeld, Marit Ackermann, Eleni G. Christodoulou, et al. "Assessing Computational Methods for Transcription Factor Target Gene Identification Based on ChIP-seq Data". In: *PLoS Computational Biology* 9.11 (2013). Ed. by Ivan Ovcharenko, e1003342.

[304] Bin Liu, Jimmy Yi, Aishwarya Sv, et al. "QChIPat: a quantitative method to identify distinct binding patterns for two biological ChIP-seq samples in different experimental conditions." In: *BMC genomics* 14 Suppl 8.Suppl 8 (2013), S3.

[305] Yuanyuan Li, David M Umbach, and Leping Li. "T-KDE: a method for genome-wide identification of constitutive protein binding sites from multiple ChIP-seq data sets." In: *BMC genomics* 15.27 (2014), p. 27.

[306] Xin Zeng, Rajendran Sanalkumar, Emery H Bresnick, et al. "jMOSAiCS: joint analysis of multiple ChIP-seq datasets." In: *Genome biology* 14.4 (2013), R38.

[307] Yajie Yang, Justin Fear, Jianhong Hu, et al. "Leveraging Biological Replicates To Improve Analysis in Chip-Seq Experiments". In: *Computational and Structural Biotechnology Journal* 9.13 (2014), pp. 1–10.

[308] Christina Schweikert, Stuart Brown, Zuojian Tang, et al. "Combining multiple ChIP-seq peak detection systems using combinatorial fusion." In: *BMC genomics* 13 Suppl 8.Suppl 8 (2012), S12.

[309] Geetu Tuteja, Peter White, Jonathan Schug, et al. "Extracting transcription factor targets from ChIP-Seq data". In: *Nucleic Acids Research* 37.17 (2009), e113.

[310] ENCODE. *TF ChIP-seq peak calling using the Irreproducibility Discovery Rate (IDR) framework*. 2012.

[311] Stan J J Brouns, Matthijs M. Jore, Magnus Lundgren, et al. "Small Crispr Rnas Guide Antiviral Defense in Prokaryotes". In: *Cancer Epidemiology Biomarkers and Prevention* 2.6 (1993), pp. 531–535. arXiv: 20.

[312] Xin Feng, Robert Grossman, and Lincoln Stein. "PeakRanger: a cloud-enabled peak caller for ChIP-seq data." In: *BMC bioinformatics* 12.1 (2011), p. 139.

[313] Han Xu, Lusy Handoko, Xueliang Wei, et al. "A signal-noise model for significance analysis of ChIP-seq with negative control". In: *Bioinformatics* 26.9 (2010), pp. 1199–1204.

[314] Guillaume J. Filion, Joke G. van Bemmel, Ulrich Braunschweig, et al. “Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells”. In: *Cell* 143.2 (2010), pp. 212–224. arXiv: NIHMS150003.

[315] Jason Ernst and Manolis Kellis. “Discovery and characterization of chromatin states for systematic annotation of the human genome.” In: *Nature biotechnology* 28.8 (2010), pp. 817–825.

[316] Jason Ernst and Manolis Kellis. “ChromHMM: automating chromatin-state discovery and characterization.” In: *Nature methods* 9.3 (2012), pp. 215–6. arXiv: NIHMS150003.

[317] Michael M Hoffman, Orion J Buske, Jie Wang, et al. “Unsupervised pattern discovery in human chromatin structure through genomic segmentation.” In: *Nature methods* 9.5 (2012), pp. 473–6.

[318] Julia Lasserre, Ho Ryun Chung, and Martin Vingron. “Finding Associations among Histone Modifications Using Sparse Partial Correlation Networks”. In: *PLoS Computational Biology* 9.9 (2013). Ed. by Niko Beerenwinkel, e1003168.

[319] Benedikt Zacher, Michael Lidschreiber, Patrick Cramer, et al. “Annotation of genomics data using bidirectional hidden Markov models unveils variations in Pol II transcription cycle. TL - 10”. In: *Molecular systems biology* 10 VN - r.12 (2014), p. 768.

[320] Jimin Song and Kevin C Chen. “Spectacle: fast chromatin state annotation using spectral learning.” In: *Genome biology* 16.1 (2015), p. 33.

[321] Alessandro Mammana and Ho-Ryun Chung. “Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome”. In: *Genome Biology* 16.1 (2015), p. 151.

[322] Hideaki Shimazaki and Shigeru Shinomoto. “A method for selecting the bin size of a time histogram.” In: *Neural computation* 19.6 (2007), pp. 1503–1527.

[323] Signal Developers. {*Signal*}*: Signal processing R package*. 2013.

[324] Gilles Celeux and Gérard Govaert. “Gaussian parsimonious mixture models”. In: *Pattern Recognition* 28.5 (1995), pp. 781–793.

[325] C Fraley, A E Raftery, M Brendan, et al. *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*. 2012.

[326] Yoav Benjamini and Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal Statistical Society B* 57.1 (1995), pp. 289–300. arXiv: 95/57289 [0035-9246].

[327] Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. “FIMO: Scanning for occurrences of a given motif”. In: *Bioinformatics* 27.7 (2011), pp. 1017–1018. arXiv: PMC3065696.

[328] Molly Megraw, Fernando Pereira, Shane T. Jensen, et al. “A transcription factor affinity-based code for mammalian transcription initiation”. In: *Genome Research* 19.4 (2009), pp. 644–656.

[329] Morten Beck Rye, Pål Sætrom, and Finn Drabløs. “A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs”. In: *Nucleic Acids Research* 39.4 (2011), e25.

[330] ENCODE Project Consortium, Bradley E Bernstein, Ewan Birney, et al. "An integrated encyclopedia of DNA elements in the human genome." In: *Nature* 489.7414 (2012), pp. 57–74. arXiv: 1111.6189v1.

[331] Qunhua Li, James B. Brown, Haiyan Huang, et al. "Measuring reproducibility of high-throughput experiments". In: *Annals of Applied Statistics* 5.3 (2011), pp. 1752–1779. arXiv: 1110.4705.

[332] H Li, B Handsaker, A Wysoker, et al. "{T}he {S}equence {A}lignment/{M}ap format and {S}{A}{M}tools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[333] Aaron R. Quinlan and Ira M. Hall. "BEDTools: A flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6 (2010), pp. 841–842.

[334] Dominic Schmidt, Petra C. Schwalie, Michael D. Wilson, et al. "Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages". In: *Cell* 148.1-2 (2012), pp. 335–348.

[335] Fidel Ramirez, Devon P Ryan, Bjorn Gruning, et al. "deepTools2: a next generation web server for deep-sequencing data analysis." In: *Nucleic acids research* 44.April (2016), pp. 160–165.

[336] Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov. "Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration". In: *Briefings in Bioinformatics* 14.2 (2013), pp. 178–192.

[337] Sascha H C Duttke, Scott A. Lacadie, Mahmoud M. Ibrahim, et al. "Perspectives on Unidirectional versus Divergent Transcription". In: *Molecular Cell* 60.3 (2015), pp. 348–349.

[338] Richard M. Myers, John Stamatoyannopoulos, Michael Snyder, et al. "A user's guide to the Encyclopedia of DNA elements (ENCODE)". In: *PLoS Biology* 9.4 (2011), e1001046.

[339] Adam M. Szalkowski and Christoph D. Schmid. "Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts". In: *Briefings in Bioinformatics* 12.6 (2011), pp. 626–633.

[340] Korbinian Strimmer. "fdrtool: A versatile R package for estimating local and tail area-based false discovery rates". In: *Bioinformatics* 24.12 (2008), pp. 1461–1462.

[341] Aaron Diaz, Kiyoub Park, Daniel A Lim, et al. "Normalization, bias correction, and peak calling for ChIP-seq." In: *Statistical applications in genetics and molecular biology* 11.3 (2012), Article 9.

[342] Kenneth J. Evans, Ni Huang, Przemyslaw Stempor, et al. "Stable Caenorhabditis elegans chromatin domains separate broadly expressed and developmentally regulated genes". In: *Proceedings of the National Academy of Sciences* 113.45 (2016), E7020–E7029.

[343] C F Beck and R A Warren. "Divergent promoters, a common form of gene organization." In: *Microbiological reviews* 52.3 (1988), pp. 318–26.

[344] Nathan T. Trinklein, Shelley Force Aldred, Sara J. Hartman, et al. "An abundance of bidirectional promoters in the human genome". In: *Genome Research* 14.1 (2004), pp. 62–66.

[345] Noritaka Adachi and Michael R. Lieber. “Bidirectional gene organization: A common architectural feature of the human genome”. In: *Cell* 109.7 (2002), pp. 807–809.

[346] Mary Q. Yang, Laura M. Koehly, and Laura L. Elnitski. “Comprehensive annotation of bidirectional promoters identifies co-regulation among breast and ovarian cancer genes”. In: *PLoS Computational Biology* 3.4 (2007), pp. 733–742.

[347] Yong H Woo and Wen-Hsiung Li. “Gene clustering pattern, promoter architecture, and gene expression stability in eukaryotic genomes.” In: *Proceedings of the National Academy of Sciences of the United States of America* 108.8 (2011), pp. 3306–11.

[348] Helen Neil, Christophe Malabat, Yves D’Aubenton-Carafa, et al. “Widespread bidirectional promoters are the major source of cryptic transcripts in yeast.” In: *Nature* 457.7232 (2009), pp. 1038–42.

[349] Françoise Wyers, Mathieu Rougemaille, Gwenaël Badis, et al. “Cryptic Pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase”. In: *Cell* 121.5 (2005), pp. 725–737.

[350] Amy C Seila, J Mauro Calabrese, Stuart S Levine, et al. “Divergent transcription from active promoters.” In: *Science (New York, NY)* 322.5909 (2008), pp. 1849–1851. arXiv: NIHMS150003.

[351] Ron A J Chen, Thomas A. Down, Przemyslaw Stempor, et al. “The landscape of RNA polymerase II transcription initiation in C. elegans reveals promoter and enhancer architectures”. In: *Genome Research* 23.8 (2013), pp. 1339–1347.

[352] Leighton J. Core, Joshua J. Waterfall, Daniel A. Gilchrist, et al. “Defining the Status of RNA Polymerase at Promoters”. In: *Cell Reports* 2.4 (2012), pp. 1025–1035. arXiv: NIHMS150003.

[353] Ho Sung Rhee and B Franklin Pugh. “Genome-wide structure and organization of eukaryotic pre-initiation complexes.” In: *Nature* 483.7389 (2012), pp. 295–301. arXiv: NIHMS150003.

[354] Leighton J Core, André L Martins, Charles G Danko, et al. “Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers”. In: *Nature Genetics* 46.12 (2014), pp. 1311–20. arXiv: NIHMS150003.

[355] Benjamin S. Scruggs, Daniel A. Gilchrist, Sergei Nechaev, et al. “Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin”. In: *Molecular Cell* 58.6 (2015), pp. 1101–1112.

[356] Shane Neph, M. Scott Kuehn, Alex P. Reynolds, et al. “BEDOPS: High-performance genomic feature operations”. In: *Bioinformatics* 28.14 (2012), pp. 1919–1920.

[357] S. L. Lauritzen and N. Wermuth. “Graphical Models for Associations Between Variables, Some of which are Qualitative and Some Quantitative”. In: *The Annals of Statistics* 17.4 (1989), pp. 1916–1916.

[358] Nir Friedman, Dan Geiger, and Moises Goldszmidt. “Bayesian Network Classifiers”. In: *Machine learning* 29.2/3 (1997), pp. 131–163.

[359] Jonathan M. Mudge and Jennifer Harrow. “Creating reference gene annotation for the mouse C57BL6/J genome assembly”. In: *Mammalian Genome* 26.9-10 (2015), pp. 366–378.

[360] Alexei A Sharov, Akira Nishiyama, Yulan Piao, et al. “Responsiveness of genes to manipulation of transcription factors in ES cells is associated with histone modifications and tissue specificity”. In: *BMC Genomics* 12.1 (2011), p. 102.

[361] Matthew E. Ritchie, Belinda Phipson, Di Wu, et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic acids research* 43.7 (2015), e47.

[362] Kevin P. Murphy. “The Bayes Net Toolbox for MATLAB”. In: *COMPUTING SCIENCE AND STATISTICS* 33 (2001), p. 2001.

[363] Juliane Perner, Julia Lasserre, Sarah Kinkley, et al. “Inference of interactions between Chromatin modifiers and histone modifications: From ChIP-Seq data to chromatin-signaling”. In: *Nucleic Acids Research* 42.22 (2014), pp. 13689–13695.

[364] Steffen L. Lauritzen and Frank Jensen. “Stable local computation with conditional Gaussian distributions”. In: *Statistics and Computing* 11.2 (2001), pp. 191–203.

[365] Robert G Cowell. “Local propagation in conditional gaussian bayesian networks”. In: *The Journal of Machine Learning Research* 6 (2005), pp. 1517–1550.

[366] Orly L. Wapinski, Thomas Vierbuchen, Kun Qu, et al. “Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons”. In: *Cell* 155.3 (2013), pp. 621–635.

[367] Soo Kyung Lee and Samuel L. Pfaff. “Synchronization of neurogenesis and motor neuron specification by direct coupling of bHLH and homeodomain transcription factors”. In: *Neuron* 38.5 (2003), pp. 731–745.

[368] Soo-Kyung Lee, Linda W Jurata, Junichi Funahashi, et al. “Analysis of embryonic motoneuron gene regulation: derepression of general activators function in concert with enhancer factors”. In: *Development (Cambridge, England)* 131.14 (2004), pp. 3295–3306.

[369] Sebastian Luehr, Holger Hartmann, and Johannes Söding. “The XXmotif web server for eXhaustive, weight matriX-based motif discovery in nucleotide sequences”. In: *Nucleic Acids Research* 40.W1 (2012), W104–9.

[370] Shaun Mahony, Matthew D. Edwards, Esteban O. Mazzoni, et al. “An integrated model of multiple-condition ChIP-Seq data reveals predeterminants of Cdx2 binding”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8394 LNBI.3 (2014). Ed. by Ilya Ioshikhes, pp. 175–176.

[371] Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, et al. “Quantifying similarity between motifs.” In: *Genome biology* 8.2 (2007), R24.

[372] U Lerner, Eran Segal, and D Koller. “Exact Inference in Networks with Discrete Children of Continuous Parents”. In: *Proceedings* (2001), pp. 238–319. arXiv: 1301.2289.

[373] S.-J. Dunn, G Martello, B Yordanov, et al. “Defining an essential transcription factor program for naive pluripotency”. In: *Science* 344.6188 (2014), pp. 1156–1160.

[374] Jonghwan Kim, Jianlin Chu, Xiaohua Shen, et al. “An Extended Transcriptional Network for Pluripotency of Embryonic Stem Cells”. In: *Cell* 132.6 (2008), pp. 1049–1061.

[375] J Wang, S Rao, J Chu, et al. “A protein interaction network for pluripotency of embryonic stem cells”. In: *Nature* 444.7117 (2006), pp. 364–368.

[376] Yuin-Han Loh, Qiang Wu, Joon-Lin Chew, et al. “The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells.” In: *Nature genetics* 38.4 (2006), pp. 431–440.

[377] Radiya G. Ali, Helen M. Bellchambers, and Ruth M. Arkell. “Zinc fingers of the cerebellum (Zic): Transcription factors and co-factors”. In: *International Journal of Biochemistry and Cell Biology* 44.11 (2012), pp. 2065–2068.

[378] Jun Aruga. “The role of Zic genes in neural development”. In: *Molecular and Cellular Neuroscience* 26.2 (2004), pp. 205–221.

[379] a. Roy, C. Francius, D. L. Rousso, et al. “Onecut transcription factors act upstream of Isl1 to regulate spinal motoneuron diversification”. In: *Development* 139.17 (2012), pp. 3109–3119.

[380] Duc N.T. Nguyen, Margaret Rohrbaugh, and Zhi-Chun Lai. “The Drosophila homolog of Onecut homeodomain proteins is a neural-specific transcriptional activator with a potential role in regulating neural differentiation”. In: *Mechanisms of Development* 97.1-2 (2000), pp. 57–72.

[381] M H Farah, J M Olson, H B Sucic, et al. “Generation of neurons by transient expression of neural bHLH proteins in mammalian cells.” In: *Development (Cambridge, England)* 127.4 (2000), pp. 693–702.

[382] Marine Lacomme, Laurence Liaubet, Fabienne Pituello, et al. “NEUROG2 drives cell cycle exit of neuronal precursors by specifically repressing a subset of cyclins acting at the G1 and S phases of the cell cycle.” In: *Molecular and cellular biology* 32.13 (2012), pp. 2596–607.

[383] Seunghee Lee, Bora Lee, Jae W. Lee, et al. “Retinoid Signaling and Neurogenin2 Function Are Coupled for the Specification of Spinal Motor Neurons through a Chromatin Modifier CBP”. In: *Neuron* 62.5 (2009), pp. 641–654.

[384] Silvia Arber, Barbara Han, Monica Mendelsohn, et al. “Requirement for the homeobox gene Hb9 in the consolidation of motor neuron identity”. In: *Neuron* 23.4 (1999), pp. 659–674.

[385] Joshua Thaler, Kathleen Harrison, Kamal Sharma, et al. “Active suppression of interneuron programs within developing motor neurons revealed by analysis of homeodomain factor HB9”. In: *Neuron* 23.4 (1999), pp. 675–687.

[386] R Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2013-2016.

[387] Ollivier Taramasco, Sebastian Bauer, and Maintainer Ollivier Taramasco. “Package RHmm”. In: (2013).

[388] Maintainer Gregory R Warnes. “Package gplots”. In: (2016).

[389] Jari Oksanen, F Guillaume Blanchet, Roeland Kindt, et al. “Package vegan”. In: (2013).

[390] MATLAB. *version R2015a*. Natick, Massachusetts: The MathWorks Inc., 2015.

Appendix I

JAMM Peak Calling Accuracy Analysis

	CTCF-K562	CTCF-HeLa	NRSF-K562	MAX-K562	Avg. Normalized Score
			Motif Precision (FIMO-based)		
JAMM-I	0.85 [1]	0.85 [0.75]	0.68 [1]	ND	0.917
JAMM-P	0.85 [1]	0.84 [0.5]	0.68 [1]	ND	0.833
DFilter	0.85 [1]	0.86 [1]	0.68 [1]	ND	1
PeakZilla	0.84 [0]	0.84 [0.5]	0.66 [0.6667]	ND	0.389
BCP	0.85 [1]	0.82 [0]	0.62 [0]	ND	0.333
PeakRanger	0.85 [1]	0.85 [0.75]	0.68 [1]	ND	0.917
MACS	0.85 [1]	0.85 [0.75]	0.68 [1]	ND	0.917

Table A.1: Peak Finding Specificity using Motif Hits Recovery. ND = Not Done. Format: original metric [0-to-1 normalized metric]. Table adapted from [73]

	CTCF-K562	CTCF-HeLa	NRSF-K562	MAX-K562	Avg. Normalized Score
			Motif Likelihood (SpeakerScan-based)		
JAMM-I	9321.09 [1]	8869.1 [1]	1445.31 [0.72]	ND	0.907
JAMM-P	8921.14 [0.9]	8574.07 [0.89]	1494.7 [0.76]	ND	0.849
DFilter	8096.96 [0.69]	8329.6 [0.79]	1332.7 [0.62]	ND	0.701
PeakZilla	8046.71 [0.67]	7357.52 [0.42]	1228.03 [0.54]	ND	0.541
BCP	5428.95 [0]	6279.35 [0]	600.58 [0]	ND	0
PeakRanger	8976.24 [0.91]	7773.98 [0.58]	1400.6 [0.68]	ND	0.724
MACS	7791.04 [0.61]	7123.08 [0.33]	1773.21 [1]	ND	0.644

Table A.2: Peak Finding Specificity using Motif Log-Likelihood. ND = Not Done. Format: original metric [0-to-1 normalized metric]. Table adapted from [73]

	CTCF-K562	CTCF-HeLa	NRSF-K562	MAX-K562	Avg. Normalized Score
			Manual Curation-based Specificity		
JAMM-I	ND	ND	59 [0.51]	164 [0.82]	0.667
JAMM-P	ND	ND	46 [0.38]	126 [0.62]	0.449
DFilter	ND	ND	17 [0.08]	10 [0]	0.041
PeakZilla	ND	ND	107 [1]	197 [1]	1
BCP	ND	ND	60 [0.52]	62 [0.28]	0.399
PeakRanger	ND	ND	57 [0.49]	151 [0.75]	0.622
MACS	ND	ND	9 [0]	16 [0.03]	0.016

Table A.3: Peak Finding Specificity using Manually Curated Peaks. ND = Not Done. Format: original metric [0-to-1 normalized metric]. Manual curation scores are taken to be the number of peaks that intersected at least one positive manually-curated peak after subtracting the number of peaks that intersected exclusively manually-curated negative peak(s). Table adapted from [73]

Appendix II

Analysis of Ngn2 Binding during Motor Neuron Programming

Ngn2 is a general pro-neuronal factor that belongs to the bHLH transcription factor family targeting the "E-box" DNA sequence motif. Forced expression of Ngn2 alone is sufficient to induce neurogenesis [381]. During motor neuron differentiation in development, Ngn2 is involved in cell cycle exit [382] but also acts as an enhancer activator [383] and cooperates with Isl1 and Lhx3, interacting via the adapter protein NLI to activate Hb9 motor neuron identity factor [384, 385] by activating an Hb9 enhancer [367, 368].

When Ngn2 is induced in embryonic bodies together with the motor neuron factors Isl1 and Lhx3, Ngn2 is degraded after 12 hours of induction although its transcription induction is not stopped [160]. Therefore, Ngn2 function required during directed motor neuron programming must be carried out within the short span of approximately 12 to 18 hours. We produced ChIP-Seq for Ngn2 at 12h after induction and found that Ngn2 and Isl1/Lhx3 are co-bound in a minority of sites ([155], Silvia Velasco and Akshay Kakumanu / Mazzoni and Mahony labs), that nonetheless appears to be important (together with Isl1/Lhx3) in activation of key motor neuron enhancers (see Chapter 6).

To understand Ngn2 function during motor neuron programming[1], we classified its binding according to whether it fell in regions that were active and accessible at 0h or whether it accessed regions previously closed an inactive at 0h, based on mixture model clustering of 0h ATAC-Seq read counts at Ngn2 binding sites (Figure B.1a). Plotting H3K27ac ChIP-Seq levels at those two classes of sites reveals that, like Isl1/Lhx3 (see Chapter 7), Ngn2 binding to previously accessible regions is correlated with enhancer decommissioning, although it appears to occur at a slower pace than that observed with Isl1/Lhx3 E1 enhancer group (Figure B.1b). This difference however is explained by

[1]Text and figures in this section are largely copied directly from [155]. For methods, please also see [155]. All ChIP-Seq and ATAC-Seq were produced by Silvia Velasco (Mazzoni lab) and Antje Hirsekorn (Ohler lab) respectively [155]

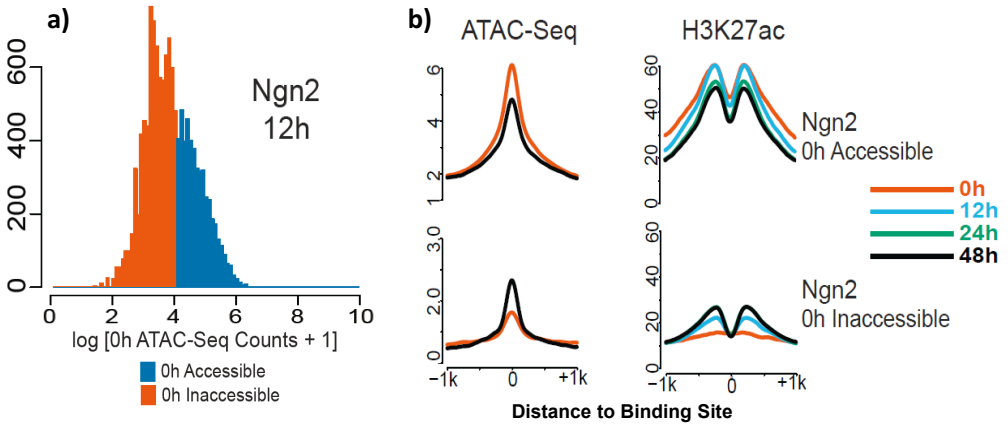


Figure B.1: a) 0h ATAC-Seq read counts at Ngn2 binding sites can be classified into two classes via mixture model clustering into accessible and inaccessible sites. b) H3K27ac levels over time during differentiation at the two different groups of sites. Figure adapted from [155].

splitting Ngn2 0h-accessible sites by whether they occur in promoters or distal regions: promoter regions active at 0h and bound by Ngn2 at 12h are not decommissioned (Figure B.2). In contrast, as expected, Ngn2 binding to previously inaccessible sites leads to sustained activation of those sites, albeit at low levels, even though Ngn2 is degraded after 12h (Figure B.1b). Indeed, Ngn2 is responsible for activation of Onecut genes responsible for activation of late acting Isl1/Lhx3 enhancers ([155], not shown).

Analysis of enrichment motifs, previously found at Isl1/Lhx3 enhancer groups, at the Ngn2 sites (Figure B3) reveals that Ngn2 binding is not enriched in any of the motifs enriched in Isl1/Lhx3 sites except, as expected, the E-box motif and a slight enrichment for Lhx motifs in Ngn2 0h-inaccessible sites consistent with enrichment of E-box motis in Isl1/Lhx3 E3 enhancer group (see Chapter 7). Therefore, consistent with the observation that Ngn2 binding is largely indepedent of Isl1/Lhx3 binding, Ngn2 does not appear to cooperate with the same transcription factors that Isl1/Lhx3 cooperate with. The cooperation between Ngn2 and Isl1/Lhx3 in a minority of sites is backed by the enrichment of Lhx3 motifs in Ngn2 binding sites. In [155], it is shown that Isl1/Lhx3 are required for those binding sites and not the opposite.

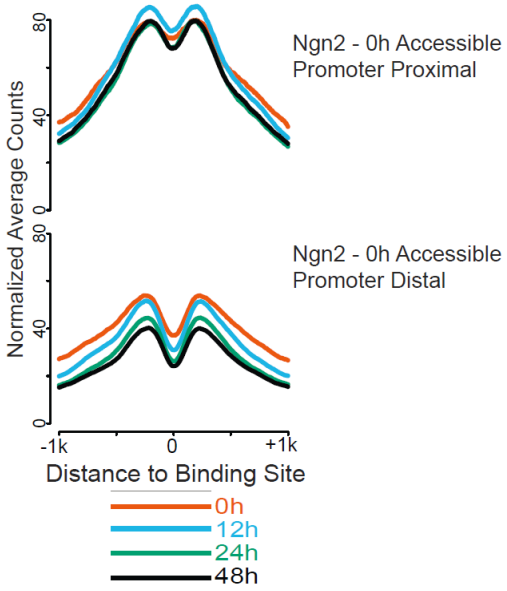


Figure B.2: H3K27ac levels over time during differentiation at Ngn2 0h-accessible proximal sites (located less than 1kb from promoter regions) and Ngn2 0h-distal accessible sites (located more than 1kb from promoters). Figure adapted from [155].

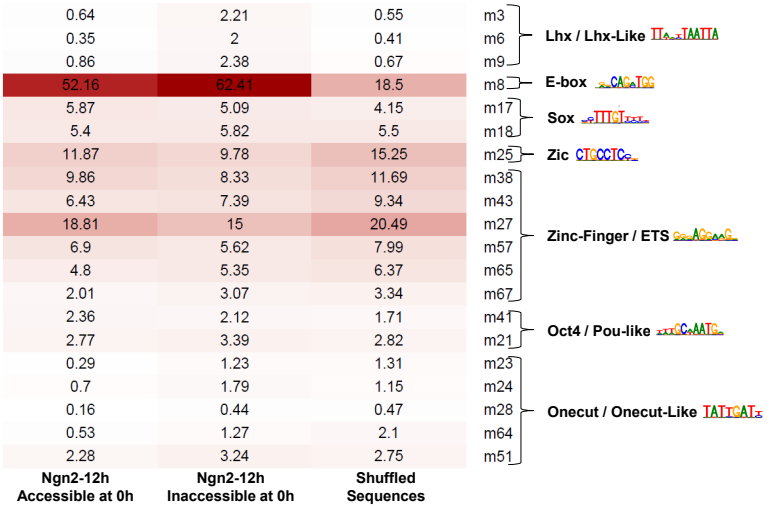


Figure B.3: Matrix shows percentage of sites where each motif occurs. Shuffled sites indicate both accessible Ngn2 and inaccessible Ngn2 sites pooled together and their sequences shuffled to maintain 2-mer frequencies.

Appendix III

List of Software and R Packages

Software described in this work

1. JAMM (first published in [73]):
   https://github.com/mahmoudibrahim/JAMM
2. hmmForChromatin (first published in [298]):
   https://github.com/mahmoudibrahim/hmmForChromatin
3. Time-course Clustering Model (first published in [155]):
   https://github.com/mahmoudibrahim/timeless

R and R packages

1. R language and environment, [386]
2. R package Signal, [323]
3. R package Mclust, [325]
4. R package RHmm, [387]
5. R package Limma, [361]
6. R package gPlots, [388]
7. R package Vegan, [389]

Third-party Software

1. bedtools, [333]
2. samtools, [332]

3. Bowtie, [58]
4. Bowtie2, [59]
5. RSEM, [63]
6. MACS, [65]
7. PeakRanger, [312]
8. PeakZilla, [70]
9. BCP, [68]
10. DFilter, [72]
11. multiGPS, [370]
12. MATLAB, [390]
13. The Bayes Net Toolbox for MATLAB, [362]